A Multi-resolution Manifold Distance for Invariant Image Similarity

Nuno Vasconcelos, *Member, IEEE,* Andrew Lippman, *Member, IEEE*

***Abstract*— Accounting for spatial image transformations is a requirement for multimedia problems such as video classification and retrieval, face/object recognition or the creation of image mosaics from video sequences. We analyze a transformation invariant metric recently proposed in the machine learning literature to measure the distance between image manifolds - the *tangent distance* (TD) - and show that it is closely related to alignment techniques from the motion analysis literature. Exposing these relationships results in benefits for the two domains. On one hand, it allows leveraging on the knowledge acquired in the alignment literature to build better classifiers. On the other, it provides a new interpretation of alignment techniques as one component of a decomposition that has interesting properties for the classification of video. In particular, we embed the TD into a multi-resolution framework that makes it significantly less prone to local minima. The new metric - *multi-resolution tangent distance* (MRTD) - can be easily combined with robust estimation procedures, and exhibits significantly higher invariance to image transformations than the TD and the Euclidean distance (ED). For classification, this translates into significant improvements in face recognition accuracy. For video characterization, it leads to a decomposition of image dissimilarity into "differences due to camera motion" plus "differences due to scene activity" that is useful for classification. Experimental results on a movie database indicate that the distance could be used as a basis for the extraction of semantic primitives such as action and romance.**

***Index Terms*— Image similarity, Manifold distance, Tangent distance, Multi-resolution, Invariance, Affine transformations, Robust estimators, Face recognition, Semantic movie classification**

I. Introduction

A large collection of problems in multimedia involve either classifying or aligning visual information. In particular, classification and alignment are a substantial component of the challenges posed by visual information retrieval and summarization. Consider the problem of finding the most similar match, in a given image database, to a query image provided by a user. This is clearly a classification problem: each image (or, if some form of labeling is available, each collection of images under the same label) in the database defines a class, and the goal is to find the class that best explains the query in the sense of minimizing the probability of retrieval error [32]. Other components of the retrieval problem, e.g. face/object detection and recognition [16], [19], [30], or extraction of semantic descriptors such as "action" vs "romance" [34] or "indoors" vs. "outdoors" [29], [31] are naturally formulated as classification problems as well. In these cases, a large collection of images of the same "theme", e.g. face or outdoors images, are assembled and used to train a classifier off-line. The classifier is then applied to the images of a particular database, labeling them with semantic tags related to that theme, e.g. "images containing people" or "images of the wild". Such labels extend the query language along semantic dimensions that greatly increase the power and usefulness of the retrieval system.

With respect to summarization, the standard solution is to segment the movie into its composing shots and select one, or a few, keyframes to represent each shot [2], [26], [37]. While this is a reasonable representation of the underlying video content, important information can be lost by completely eliminating the shot's dynamic component. For example, it may become impossible to distinguish two shots of a movie where the same people perform different actions on the same set. A better sense of the scene dynamics is achieved through a mosaic [10], [14], [21] that presents the average of all the images after alignment according to the dominant motion in the scene (typically that of the camera). If the registration is precise, static objects appear crisp while moving objects create a smooth spatial trail determined by their motion. This is generally sufficient to enable a coarse understanding of object motion through the entire shot from the observation of the static mosaic. Still better rendition of the scene dynamics can be achieved with layering [21], [35], [36]. Here, each frame is segmented into the composing objects, and an individual mosaic created for each object. The combination of this mosaic with a segmentation mask for each frame and the object's motion allows the perfect reconstruction of the objects evolution in the scene.

For both retrieval and mosaic creation, significant performance improvements are achievable by relying on precise image alignment. In the case of retrieval, alignment is a means to achieve invariance against spatial transformations such as rotation or scaling. For mosaic creation, alignment is the fundamental problem since without it the resulting mosaics or layers will simply render an arbitrary average of the individual frames and will not reflect the scene or its dynamics. In fact, as we will show below, alignment and classification can be seen as two sides of the same coin: while, on one hand, the appropriate distance for classification is that which maximizes alignment, on the other, classification requires very little else once alignment is reached.

Despite these synergies, there are few unified treatments of the two problems. In the vision literature, while a significant body of work has been devoted to alignment (or recovery of

N. Vasconcelos is with the Department of Electrical and Computer Engineering, University of California, San Diego. A. Lippman is with the Media Laboratory, Massachussets Institute of Technology.

motion parameters), considerably smaller attention has been given to the question of how to explicitly account for it in the context of classification [8]. Instead, invariance is usually encoded in the features [12], [17], [22] or learned from examples [20], [23], [27]. Such solutions are not always satisfying: invariant features can be quite arbitrary and it is usually difficult to evaluate the impact on the classification error of the information that is discarded, learning has combinatorial complexity on the number of degrees of freedom of the transformations to be learned [8]. Conversely, classification has received tremendous attention in the learning literature, where little attention has been given to the problem of visual alignment.

One exception to this rule is the TD classifier introduced in [24]. The key idea behind the TD is that, when subject to spatial transformations, images span manifolds in high dimensional Euclidean space, and an invariant metric should measure the distance between those manifolds instead of the distance between other properties of (or features extracted from) the images themselves. The distance between two manifolds is defined as the ED between their closest points. Because these manifolds may have complex shapes, the resulting optimization problem is usually a difficult one. It can, nevertheless, be made tractable by considering the minimization of the distance between the manifolds' tangent spaces - the TD - instead of that between the manifolds themselves. It turns out that the tangent hyperplane to a manifold at the point corresponding to a given image, is the first-oder Taylor series expansion of the image intensity function. This expansion has been widely used in the motion analysis literature (since [11]), and is well known to hold only locally, i.e. when the ED between the images to align is small.

Making the connection between the TD classifier and image alignment techniques therefore explains one of the major limitations of the former: while leading to impressive results for the problem of character recognition [25], it cannot handle well natural images since these are usually subject to a larger set of image transformations. In this paper, we make the connection explicit by formulating recognition as classification, alignment as regression, and showing that the particular classification architecture on which the TD classifier is based, known as *nearest neighbors*, actually embeds a regression problem in the decision function used for classification. The TD classifier can, therefore, be seen as solving the alignment problem for each evaluation of the decision function. The new interpretation allows leveraging on the knowledge acquired in the alignment literature to improve the classification performance. In particular, we use the fact that, by extending the range over which linear approximations hold, multi-resolution decompositions significantly improve the performance of image registration algorithms based on the Taylor series approximation. In the context of classification, this leads to a classifier that embeds the computation of the TD on a multi-resolution framework [5]. We denote the new metric by *multi-resolution tangent distance* (MRTD) and evaluate its performance on the task of face recognition. These experiments show that, when compared to the TD or ED, the MRTD exhibits significantly higher invariance to image transformations.

From the point of view of image alignment, the connection to classification is important because it emphasizes the fact that what cannot be explained by the alignment model, the MRTD, is a significant piece of the information about two images. In fact, it leads to a decomposition into "alignment parameters plus what cannot be explained by alignment" which is, for some alignment models, interesting by itself. We illustrate this property by showing that when combined with simple models of camera motion, such as affine transformations, the TD can be interpreted as a metric of the activity in a video sequence, an important feature for the semantic characterization of a movie. Experiments on a movie database show that the simple integration of the MRTD throughout a scene is a good descriptor for the action content of that scene.

As a metric of image similarity, the MRTD is shown to have several appealing properties: 1) maintains the general purpose nature of the TD; 2) can be easily combined with robust estimation procedures, exhibiting invariance to moderate non-linear image variations (such as those caused by slight variations in shape or occlusions); 3) is amenable to computationally efficient screening techniques where bad matches are discarded at low resolutions; 4) performs well on recognition tasks; and 5) enables the design of a single architecture for problems as diverse as face recognition, semantic video classification, and mosaic creation.

II. Classification

Consider a classification problem where a query pattern $\mathbf{z}$ is to be classified into one of C classes. Both $\mathbf{z}$ and C can vary depending on the classification domain. For example, in image retrieval C is the number of image classes in the database, while $\mathbf{z}$ can be a feature (e.g. a color histogram) or collection of features (e.g. a collection of wavelet coefficients) extracted from a query image. On the other hand, for face recognition $C = 2$ ("face" and "non-face" classes) and $\mathbf{z}$ an image patch. Defining a class-indicator variable $Y \in \{1, \ldots, C\}$ and denoting by $\mathbf{Z}$ the random variable according to which the observed patterns are drawn[1], it is well know that, when the goal is minimize the probability of classification error, the optimal solution is provided by the *Bayes classifier* [6], [7]

$$g^*(\mathbf{z}) = \arg\max_i P_{Y|\mathbf{Z}}(i|\mathbf{z}). \tag{1}$$

Furthermore, the probability of error is lower bounded by the *Bayes error*

$$L^* = 1 - E_{\mathbf{z}}[\max_i P_{Y|\mathbf{Z}}(i|\mathbf{z})], \tag{2}$$

where $E_{\mathbf{z}}$ means expectation with respect to $P_{\mathbf{Z}}(\mathbf{z})$.

The Bayes classifier is not always easy to implement in practice. A simpler and very popular alternative is the nearest neighbors classifier. Denoting by $\mathbf{t}_i = \{\mathbf{t}_{i,1}, \ldots, \mathbf{t}_{i,N_i}\}$ the

[1]We use upper case for random variables and lower case for particular values, e.g. $\mathbf{Z} = \mathbf{z}$ denotes that the random variable $\mathbf{Z}$ takes the value $\mathbf{z}$. When the meaning is clear from context, we usually omit one of the symbols. For example, $P_{\mathbf{Z}|Y}(\mathbf{z}|i)$ is commonly used instead of $P_{\mathbf{Z}|Y}(\mathbf{z} = \mathbf{z}|Y = i)$. Boldface type is used to represent vectors.

training sample for the i^{th} class, it corresponds to the decision function

$$g(\mathbf{z}) = \arg\min_i \{\min_j \mathcal{D}(\mathbf{z}, \mathbf{t}_{i,j})\}, \quad (3)$$

where $\mathcal{D}$ is a metric, typically the ED. A common extension is the k-nearest neighbors classifier where the minimization above is replaced by a majority vote among the k training points that are closest to $\mathbf{z}$. In addition to its simplicity, kNN rules are attractive because it can be shown that their probability of error is upper bounded by [6]

$$L_{kNN} \leq L^*(1 + O(\frac{1}{\sqrt{k}})), \quad (4)$$

where L^* is the Bayes error. Hence, even for $k = 1$, there is a guarantee that the probability of error will be at most twice the Bayes error. Even though all that is presented in this work is valid for k-nearest neighbor classifiers, for simplicity we concentrate on the nearest neighbor case.

III. Regression

Regression is a statistical technique for modeling relations between variables [15]. The most popular regression model is

$$\mathbf{z} = f_{\mathbf{p}}(\mathbf{t}) + \epsilon, \quad (5)$$

where $\mathbf{t}$ is a *predictor* variable, $\mathbf{z}$ a *response* variable, and ϵ a random variable that accounts for the noise associated with the observation of the response variable. The function $f_{\mathbf{p}}$ belongs to a family of functions parameterized by the *parameter vector* $\mathbf{p}$. Other regression models are possible, e.g. models that allow noise not only in the observation of the response but also on the predictor itself, but we will not consider them here.

Given a probability density for the noise $P_\epsilon(\epsilon)$ it is straightforward to see, from (5), that

$$P_{\mathbf{Z}|\mathbf{T}}(\mathbf{z}|\mathbf{t}) = P_\epsilon(\mathbf{z} - f_{\mathbf{p}}(\mathbf{t})). \quad (6)$$

The goal is, for a given training sequence of observed pairs (t_i, z_i), to find the parameter vector that maximizes the likelihood of the observations under this model

$$\mathbf{p}^* = \arg\max_{\mathbf{p}} P_\epsilon(\mathbf{z} - f_{\mathbf{p}}(\mathbf{t})), \quad (7)$$

where $\mathbf{z}$ and $\mathbf{t}$ are the vectors with entries z_i and t_i, respectively. A common assumption is that the noise is a a zero-mean stochastic process from the exponential family

$$P_\epsilon(\epsilon) = Ke^{-\mathcal{F}(\epsilon)} \quad (8)$$

where K is a normalizing constant. In this case, (7) reduces to

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \mathcal{F}(\mathbf{z} - f_{\mathbf{p}}(\mathbf{t})) \quad (9)$$

which can usually be rewritten as

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \mathcal{D}(\mathbf{z}, f_{\mathbf{p}}(\mathbf{t})), \quad (10)$$

where $\mathcal{D}$ is a metric. For example, when the noise samples are independently distributed and Gaussian, $\mathcal{D}$ is the ED

$$\mathcal{D}(\mathbf{z}, f_{\mathbf{p}}(\mathbf{t})) = ||\mathbf{z} - f_{\mathbf{p}}(\mathbf{t})||^2 = \sum_i (z_i - [f_{\mathbf{p}}(\mathbf{t})]_i)^2. \quad (11)$$

The problem of image alignment is naturally formulated as a regression problem. Consider two image patches with pixel intensities $M(\mathbf{x}_i)$ and $N(\mathbf{x}_i)$, where $\mathbf{x}_i$ is the 2D vector of image coordinates of pixel i, and the manifold spanned by all the possible spatial transformations

$$T_{\mathbf{p}}[M(\mathbf{x}_i)] = M(\psi(\mathbf{x}_i, \mathbf{p})), \quad (12)$$

that a pattern may be subject to, where ψ is a function (typically) linear on $\mathbf{p}$, but not necessarily linear on $\mathbf{x}_i$. Letting $z_i = M(\mathbf{x}_i)$, $t_i = N(\mathbf{x}_i)$, and $f_{\mathbf{p}}(t_i) = T_{\mathbf{p}}[N(\mathbf{x}_i)]$, the optimal parameter vector is

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \mathcal{D}(M(\mathbf{x}), T_{\mathbf{p}}[N(\mathbf{x})]). \quad (13)$$

IV. The tangent distance

Comparing (10) with (3) it can be seen that, for each i and j, the standard nearest neighbor classifier solves a regression problem. The particular aspect of this regression problem is that the family of functions $f_{\mathbf{p}}(\mathbf{t})$ is restricted to the identity map, leading to the trivial solution $f_{\mathbf{p}^*}(\mathbf{t}) = \mathbf{t}$. The TD classifier relaxes this constraint by allowing a generic regression problem inside the decision function (3).

The main idea is that the distances in which the classification is based should be those between the manifolds spanned by the query pattern and that in the training set, not the distances between the patterns themselves. This is illustrated in Figure 1.

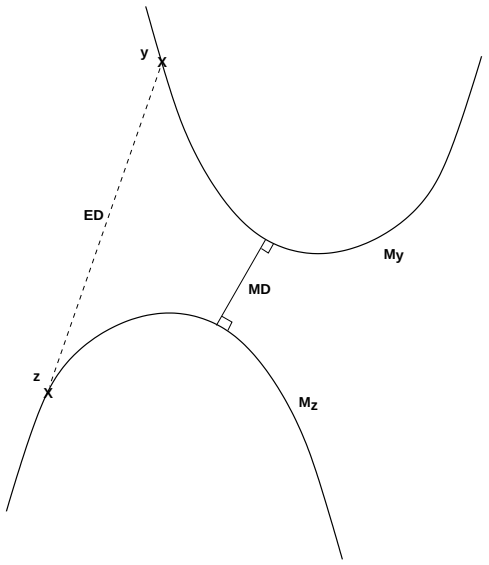

Fig. 1. The Euclidean distance (ED) between patterns $\mathbf{y}$ and $\mathbf{z}$, and manifold distance (MD) between the corresponding manifolds $\mathbf{M_y}$ and $\mathbf{M_z}$.

Given two patterns $M(\mathbf{x})$ and $N(\mathbf{x})$, the distance between the associated manifolds - *manifold distance* (MD) - is

$$\mathcal{T}(M, N) = \min_{\mathbf{p},\mathbf{q}} ||T_{\mathbf{q}}[M(\mathbf{x})] - T_{\mathbf{p}}[N(\mathbf{x})]||^2, \quad (14)$$

where $||.||$ is the Euclidean norm (11). For simplicity, we consider a version of the distance in which only one of the patterns is subject to a transformation, i.e.

$$\mathcal{T}(M, N) = \min_{\mathbf{p}} ||M(\mathbf{x}) - T_{\mathbf{p}}[N(\mathbf{x})]||^2, \quad (15)$$

but all results can be extended to the two-sided distance. Notice that this is exactly the image alignment equation of (13) (when $\mathcal{D}$ is the Euclidean norm).

Since the pixel intensity $N(\mathbf{x})$ is usually a highly nonlinear function of the image coordinates, there is, in general, no closed form solution for (15). A well known trick from the alignment literature is to linearize $T_{\mathbf{p}}[N(\mathbf{x})]$ through a first order Taylor series expansion [3], [11]. Using the fact that

$$\begin{aligned}\nabla_{\mathbf{p}} T_{\mathbf{p}}[N(\mathbf{x})] &= \nabla_{\mathbf{p}} N(\psi(\mathbf{x},\mathbf{p})) \\ &= \nabla_{\mathbf{p}}\psi(\mathbf{x},\mathbf{p})\nabla_{\mathbf{x}} N(\psi(\mathbf{x},\mathbf{p})), \end{aligned} \quad (16)$$

where $\nabla_{\mathbf{p}} T_{\mathbf{p}}$ is the gradient of $T_{\mathbf{p}}$ with respect to $\mathbf{p}$, $T_{\mathbf{p}}[N(\mathbf{x})]$ can, for small $\mathbf{p}$, be approximated by a first order Taylor expansion around the identity transformation

$$T_{\mathbf{p}}[N(\mathbf{x})] = N(\mathbf{x}) + (\mathbf{p}-\mathbf{I})^T \nabla_{\mathbf{p}}\psi(\mathbf{x},\mathbf{p})\nabla_{\mathbf{x}} N(\mathbf{x}).$$

As shown in [24], this is equivalent to approximating the manifold by a tangent hyper-plane, and leads to the TD. Substituting this expression in (15), setting the gradient with respect to $\mathbf{p}$ to zero, and solving for $\mathbf{p}$ leads to

$$\begin{aligned}\mathbf{p} &= \left[\sum_{\mathbf{x}} \nabla_{\mathbf{p}}\psi(\mathbf{x},\mathbf{p})\nabla_{\mathbf{x}} N(\mathbf{x})\nabla_{\mathbf{x}}^T N(\mathbf{x})\nabla_{\mathbf{p}}^T\psi(\mathbf{x},\mathbf{p})\right]^{-1} \\ &\quad \times \sum_{\mathbf{x}} D(\mathbf{x})\nabla_{\mathbf{p}}\psi(\mathbf{x},\mathbf{p})\nabla_{\mathbf{x}} N(\mathbf{x}) + \mathbf{I}, \end{aligned} \quad (17)$$

where $D(\mathbf{x}) = M(\mathbf{x}) - N(\mathbf{x})$. Given this optimal $\mathbf{p}$, the TD between the two patterns is computed with (12) and (15). The main limitation of this formulation is that, since it is a first-order approximation, it is only valid for a small range of variation in the parameter vector $\mathbf{p}$. This is illustrated in Figure 2.

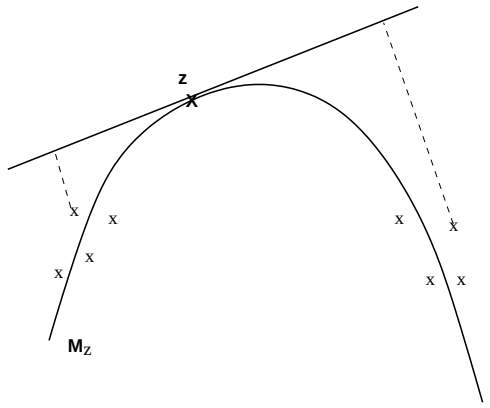

Fig. 2. Outside a narrow range of transformations, points close to the manifold $\mathbf{M_z}$ can have large TD (shown as a dashed line). Notice that for the points on the right, the distance between the TD and the MD is much larger than that for those on the left. I.e. the error depends on how the manifold deviates from a plane.

A. Manifold distance via Newton's method

As an alternative to linearization, the minimization of the MD (15) can be performed through iterative procedures such as Newton's method

$$\mathbf{p}^{n+1} = \mathbf{p}^n - \alpha \left[\nabla_{\mathbf{p}}^2 \mathcal{T}|_{\mathbf{p}=\mathbf{p}^n}\right]^{-1} \nabla_{\mathbf{p}} \mathcal{T}|_{\mathbf{p}=\mathbf{p}^n}, \quad (18)$$

where $\nabla_{\mathbf{p}}\mathcal{T}$ and $\nabla_{\mathbf{p}}^2\mathcal{T}$ are, respectively, the gradient and Hessian of the cost function (15) with respect to the parameter $\mathbf{p}$,

$$\nabla_{\mathbf{p}}\mathcal{T} = 2\sum_{\mathbf{x}} \left[M(\mathbf{x}) - T_{\mathbf{p}}[N(\mathbf{x})]\right] \nabla_{\mathbf{p}} T_{\mathbf{p}}[N(\mathbf{x})]$$

$$\begin{aligned}\nabla_{\mathbf{p}}^2\mathcal{T} &= 2\sum_{\mathbf{x}} \left[-\nabla_{\mathbf{p}} T_{\mathbf{p}}[N(\mathbf{x})]\, \nabla_{\mathbf{p}}^T T_{\mathbf{p}}[N(\mathbf{x})] + \right. \\ &\quad \left. + \left[M(\mathbf{x}) - N(\mathbf{x})\right] \nabla_{\mathbf{p}}^2 T_{\mathbf{p}}[N(\mathbf{x})]\right] \end{aligned}$$

and $\mathbf{p}^n$ the optimal solution at iteration n.

Disregarding second-order order terms, choosing $\mathbf{p}^0 = \mathbf{I}$ and $\alpha = 1$, using (16), and substituting in (18) leads to (17). I.e. the TD corresponds to a single iteration of the minimization of the MD by a simplified version of Newton's method, where second-order derivatives are disregarded. This reduces the rate of convergence of Newton's method, and a single iteration may not be enough to achieve the local minimum, even for simple functions. It is, therefore, possible to achieve improvement if iteration (18) is repeated until convergence.

V. Extensions to the TD

The iterative minimization of (18) suffers from two major drawbacks [4]: 1) it may require a significant number of iterations for convergence, and 2) it can easily get trapped in local minima. Both these limitations can be at least partially avoided by embedding the computation of the MD in a multi-resolution framework, leading to the *multi-resolution manifold distance* (MRMD).

A. The multi-resolution manifold distance

To compute the MRMD, the patterns to classify are first subject to a multi-resolution decomposition (such as a Gaussian pyramid [5]), and the MD is then iteratively computed for each layer, using the estimate obtained from the layer above as a starting point,

$$\begin{aligned}\mathbf{p}_l^{n+1} &= \mathbf{p}_l^n + \alpha \left[\sum_{\mathbf{x}} \nabla_{\mathbf{p}} T_{\mathbf{p}_l^n}[N(\mathbf{x})] \nabla_{\mathbf{p}}^T T_{\mathbf{p}_l^n}[N(\mathbf{x})]\right]^{-1} \\ &\quad \times \sum_{\mathbf{x}} D_l^n(\mathbf{x}) \nabla_{\mathbf{p}} T_{\mathbf{p}_l^n}[N(\mathbf{x})], \end{aligned} \quad (19)$$

where, $D_l^n(\mathbf{x}) = M(\mathbf{x}) - T_{\mathbf{p}_l^n}[N(\mathbf{x})]$. If only one iteration is allowed at each image resolution, the MRMD becomes the multi-resolution extension of the TD, i.e. the *multi-resolution tangent distance* (MRTD).

To illustrate the benefits of minimization over different scales consider the signal

$$f(t) = \sum_{k=1}^{K} \sin(w_k t),$$

consisting of a sum of sinusoids at frequencies w_k which are multiples of a fundamental frequency $w_k = k w_0, k = 1,\ldots,K$, and the manifold generated by all its possible translations

$$f'(t,d) = f(t+d) = \sum_{k=1}^{K} \sin(w_k (t+d)).$$

For a given translation d, the ED between the original and translated functions is

$$D(d) = \int \left[\sum_{k=1}^{K} (\sin(w_k t) - \sin(w_k (t+d))) \right]^2 dt,$$

and the corresponding manifold distance

$$\mathcal{T} = \min_d D(d).$$

Figure 3 depicts the multi-resolution Gaussian decomposition of $f(t)$, together with the ED between $f(t)$ and $f'(t,d)$ as a function of the translation d. Notice that as the resolution increases, the distance function has more local minima, indicating that the manifold is "bumpier". Therefore, even when the patterns to align are on the manifold, the range of translations with guaranteed convergence to the global minimum (at $d = 0$) decreases inversely to the resolution. I.e., at higher resolutions, a better initial estimate is necessary to obtain the same performance from the minimization algorithm.

Notice also that, since the function to minimize is very smooth at the lowest resolutions, the minimization will require few iterations at these resolutions if a procedure such as Newton's method is employed. Furthermore, since the minimum at one resolution is a good guess for the minimum at the next resolution, the computational effort required to reach that minimum will also be small. Finally, since a minimum at low resolutions is based on coarse, or global, information about the function or patterns to be classified, it is likely to be the global minimum of at least a significant region of the parameter space, if not the true global minimum.

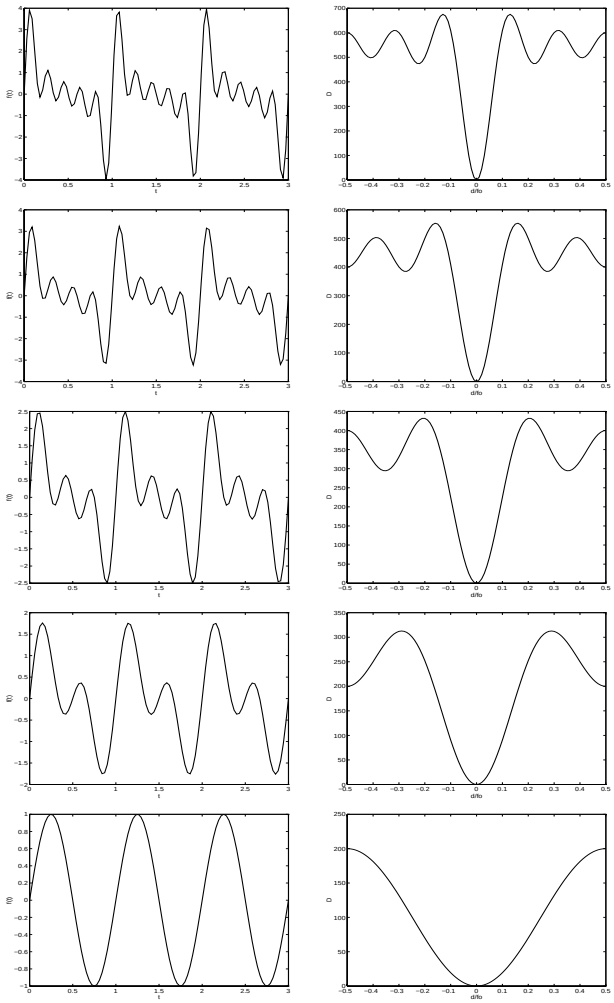

Fig. 3. Left: Five scales of the multi-resolution decomposition of $f(t)$. Right: Euclidean distance vs. translation for each scale. Resolution decreases from top to bottom.

B. Affine-invariant classification

There are many linear transformations that can be used in (12). In this work, we consider manifolds generated by affine transformations

$$\psi(\mathbf{x}, \mathbf{p}) = \begin{bmatrix} x & y & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x & y & 1 \end{bmatrix} \mathbf{p} = \mathbf{\Phi}(\mathbf{x})\mathbf{p}, \qquad (20)$$

where $\mathbf{p} = (p_{xx}, p_{xy}, p_{x0}, p_{yx}, p_{yy}, p_{y0})^T$ is the vector of parameters which characterize the transformation. Taking the gradient of (20) with respect to $\mathbf{p}$, $\nabla_{\mathbf{p}} \psi(\mathbf{x}, \mathbf{p}) = \mathbf{\Phi}(\mathbf{x})^T$, using (16), and substituting in (19),

$$\begin{aligned} \mathbf{p}_l^{n+1} &= \mathbf{p}_l^n + \\ &\alpha \left[\sum_{\mathbf{x}} \mathbf{\Phi}(\mathbf{x})^T \nabla_{\mathbf{x}} N'(\mathbf{x}) \nabla_{\mathbf{x}}^T N'(\mathbf{x}) \mathbf{\Phi}(\mathbf{x})^T \right]^{-1} \\ &\times \sum_{\mathbf{x}} D'(\mathbf{x}) \mathbf{\Phi}(\mathbf{x})^T \nabla_{\mathbf{x}} N'(\mathbf{x}), \end{aligned} \qquad (21)$$

where $N'(\mathbf{x}) = N(\psi(\mathbf{x}, \mathbf{p}_l^n))$, and $D'(\mathbf{x}) = M(\mathbf{x}) - N'(\mathbf{x})$. For a given level l of the multi-resolution decomposition, the iterative process of (21) can be summarized as follows.

1) Compute $N'(\mathbf{x})$ by warping the pattern to classify $N(\mathbf{x})$ according to the best current estimate of $\mathbf{p}$, and compute its spatial gradient $\nabla_{\mathbf{x}} N'(\mathbf{x})$.
2) Update the estimate of $\mathbf{p}_l$ according to (21).
3) Stop if convergence, otherwise go to 1.

The parameter α must be found through a line search [4] in order to guarantee a decrease of the cost function at each iteration. The simplest way to achieve this is to consider a sequence $\alpha_i = \alpha_{i-1}^2$, with $\alpha_0 = 1$. These α_i are successively tried in step 2, until the ED between $M(\mathbf{x})$ and $N(\mathbf{x})$ warped according to $\mathbf{p}_l^{n+1}$ is smaller than that obtained with $\mathbf{p}_l^n$. In practice, it suffices to try two or three values of α since a very small α indicates convergence. Once the final $\mathbf{p}_l$ is obtained, it is passed to the multi-resolution level below (by doubling the translation parameters), where it is used as initial estimate. Since the initial guess provided by the higher level of the pyramid is, in general, close to the actual minimum, the iterative procedure of steps 1-3 usually converges within a small number of iterations. Given the values of $\mathbf{p}_i$ that minimize the MD between a pattern to classify and a set of prototypes in the database, a K-nearest neighbor classifier is used to find the pattern's class.

C. Robust classifiers

One issue of importance for classification systems is that of robustness to outliers, i.e errors that occur with low probability,

but which can have large magnitude. Examples are errors due to variation of facial features (e.g. faces shot with or without glasses) in face recognition, errors due to undesired blobs of ink or uneven line thickness in character recognition, or errors due to partial occlusions (such as a hand in front of a face) or partially missing patterns (such as an undoted *i*). It is well known that a few (maybe even one) outliers of high leverage are sufficient to throw mean squared error estimators completely off-track [18].

Several robust estimators have been proposed in the statistics literature to avoid this problem. In this work we consider *M-estimators* [9] which can be very easily incorporated in the MD classification framework. M-estimators are an extension of least squares estimators where the square function is substituted by a functional $\rho(x)$ which weighs large errors less heavily. The robust-estimator version of the MD then becomes to minimize the cost function

$$\mathcal{T}(M,N) = \min_{\mathbf{p}} \sum_{\mathbf{x}} \rho(M(\mathbf{x}) - T_{\mathbf{p}}[N(\mathbf{x})]), \qquad (22)$$

and it is straightforward to show that the "robust" equivalent to (21) is

$$\mathbf{p}_l^{n+1} = \mathbf{p}_l^n + \alpha \left[\sum_{\mathbf{x}} \rho''[D(\mathbf{x})] \mathbf{\Phi}(\mathbf{x})^T \nabla_{\mathbf{x}} N'(\mathbf{x}) \nabla_{\mathbf{x}}^T N'(\mathbf{x}) \mathbf{\Phi}(\mathbf{x})^T \right]^{-1} \times \left[\sum_{\mathbf{x}} \rho'[D(\mathbf{x})] \mathbf{\Phi}(\mathbf{x})^T \nabla_{\mathbf{x}} N'(\mathbf{x}) \right], \qquad (23)$$

where $D(\mathbf{x}) = M(\mathbf{x}) - N'(\mathbf{x})$ and $\rho'(x)$ and $\rho''(x)$ are, respectively, the first and second derivatives of the function $\rho(x)$ with respect to its argument.

VI. Experimental evaluation

In this section, we report on experiments carried out to evaluate the performance of the MRTD in various tasks. The first set of experiments was designed to illustrate the invariance of the TD to affine transformations, and compare the range of invariance attained with the different extensions. The second set demonstrates the benefits of the MRTD for a classification task: face recognition. Finally, the third set illustrates how the intuitive interpretation of the TD as "all that cannot be explained by affine transformations" can be useful for important multimedia applications such as the semantic classification of movies.

A. Affine invariance of the TD

Starting from a single view of a reference image, we created an artificial dataset composed by 441 affine transformations of it. These transformations consisted of combinations of all rotations in the range from -30 to 30 degrees with increments of 3 degrees, with all scaling transformations in the range from 70% to 130% with increments of 3%. The images associated with the extremes of the scaling/rotation space are represented on the top portion of Figure 4.

On the bottom of Figure 4 are the distance surfaces obtained by measuring the distance associated with several metrics at each of the points in the scaling/rotation space. Five metrics were considered in this experiment: ED, the TD, the MD computed through Newton's method, the MRMD, and the MRTD. While the TD exhibits some invariance to rotation and scaling, this invariance is restricted to a small range of the parameter space and performance only slightly better than the obtained with the ED. The performance of the MD computed through Newton's method is dramatically superior, but still inferior to those achieved with the MRTD (which is very close to zero over the entire parameter space considered in this experiment), and the MRMD. The performance of the MRTD is in fact impressive given that it involves a computational increase of less than 50% with respect to the TD, while each iteration of Newton's method requires an increase of 100%, and several iterations are typically necessary to attain the minimum MD.

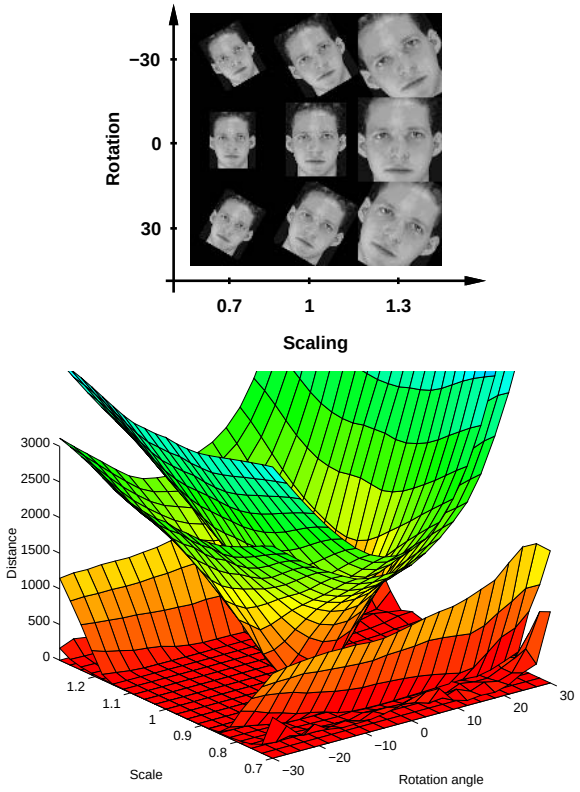

Fig. 4. Invariance of the tangent distance. In the bottom picture, the surfaces shown correspond to ED, TD, MD through Newton's method, MRTD, and MRMD. This ordering corresponds to that of the nesting of the surfaces, i.e. the ED is the cup-shaped surface in the center, while the MRMD is the flat surface which is approximately zero everywhere.

B. Face recognition

To evaluate the performance of the MRTD on a classification task, we conducted a series of face recognition experiments, using the Olivetti Research Laboratories (ORL) face database. This database is composed by 400 images of 40 subjects, 10 images per subject, and contains some variation in pose, illumination, expressions and facial features. On the other hand it exhibits almost no variation in terms of scaling, or in-plane head rotation, and assumes no translation, i.e. all faces are centered at approximately the same position.

Fig. 5. A subset of the ORL face database.

For these reasons, the ORL database is a suitable candidate for the controlled experiments required to quantify the impact on the recognition accuracy of the different degrees of invariance achieved by different extensions of the TD. The idea is to start from the original set of faces in the canonical pose (a small subset of which is presented in Figure 5) and create several replicas by applying different degrees of translation, scaling and rotation. The dependence of recognition accuracy on these variables can then be quantified by simply measuring the recognition rates on each dataset. We created three artificial datasets by applying to each image a random transformation drawn from a multivariate normal distribution with zero average displacement and rotation, unitary average scaling and the standard deviations presented in Table I. They are ordered by increasing variability, i.e. degree of difficulty that they pose to the recognition task. Figure 6 presents the samples corresponding to that of Figure 5 for each of the three new datasets.

Dataset	σ_x	σ_y	σ_r	σ_s
D1	4	3	5	1
D2	4	3	10	5
D3	4	3	20	10

TABLE I

STANDARD DEVIATION OF THE MULTIVARIATE NORMAL DENSITIES FROM WHICH THE IMAGE TRANSFORMATIONS WERE DRAWN. σ_x, AND σ_y REFER TO TRANSLATION, σ_r TO DEGREES OF ROTATION, AND σ_s TO PERCENT SCALING.

We next designed three experiments with increasing degree of difficulty. In the first, we selected the first view of each subject as the test set, using the remaining nine views as training data. In the second, the first five faces were used as test data while the remaining five were used for training. Finally, in the third experiment, we reverted the roles of the datasets used in the first. The recognition accuracy for each of these experiments and each of the datasets is reported on Table II for the ED, the TD, the MRTD, and a robust version of this distance (RMRTD) with

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } x \leq \sigma T \\ \frac{T^2}{2}, & \text{if } x > \sigma T, \end{cases}$$

where T is a threshold (set to 2 in our experiments), and σ a robust version of the error standard deviation defined as σ = median $|e_i$ - median $(e_i)|$ / 0.6745. Notice that since for points such that $x > \sigma T$ both the first and second derivatives of the robust functional are zero, this estimator simply disregards outliers. All results were obtained with a simple nearest neighbor classifier to maintain consistency across experiments.

Fig. 6. Transformed versions of the sample of figure 5, according to the parameters of Table I. Top to bottom: datasets D1, D2, and D3.

Bar graphs corresponding to these tables are plotted in Figure 7. It is clear that the multi-resolution distances provide a significantly higher invariance to linear transformations than the ED or the TD, increasing the recognition accuracy by as much as 37.8% in the hardest dataset (D3). In fact, for the easier tasks of experiments one and two, the performance of the multi-resolution classifiers is almost constant and accuracy always above 90%. It is only for the harder experiment that their invariance starts to break down. Even in this case the degradation is graceful - recognition accuracy only drops below 75% for considerable rotation and scaling (dataset D3). On the other hand, the ED and the single resolution TD break down even for the easier tasks, and fail dramatically when the hardest task is performed on the more difficult datasets. Among the multi-resolution distances the best performance is

Train/Test	Distance	D0	D1	D2	D3
9/1 (exp. 1)	ED	97.5	82.5	75.0	60.0
	TD	97.5	92.5	85.0	70.0
	MRTD	100.0	100.0	100.0	97.5
	RMRTD	100.0	100.0	100.0	97.5
5/5 (exp. 2)	ED	92.0	81.5	82.5	82.5
	TD	92.5	88.0	85.5	84.0
	MRTD	95.0	95.0	96.0	92.5
	RMRTD	95.5	95.5	95.0	92.0
1/9 (exp. 3)	ED	71.1	39.7	35.3	21.7
	TD	73.6	49.2	40.0	24.7
	MRTD	75.0	75.6	73.0	59.7
	RMRTD	79.1	78.6	75.5	62.5

TABLE II

CLASSIFICATION ACCURACY (PERCENTAGE OF FACES CORRECTLY RECOGNIZED) FOR THE THREE EXPERIMENTS DISCUSSED IN THE TEXT. D0 IS THE DATASET OBTAINED FROM THE ORL DATABASE, WHILE D1, D2 AND D3 WERE OBTAINED THROUGH THE TRANSFORMATIONS OF TABLE I.

achieved by the RMRTD. A significant gain over the MRTD is, however, only observed in the hardest problem (D3) indicating that the MRTD is a sufficient solution whenever various examples of the faces to recognize are available for training.

C. Detailed analysis

While the previous results demonstrate that significant gains can be achieved by replacing the TD classifier by its multi-resolution counterpart, a thorough understanding of the properties of the MRTD requires additional experiments. As seen in section V the MRTD classifier propagates the parameter estimates obtained at a given resolution to obtain initial estimates at the next. Errors can therefore occur whenever the low-resolution estimates are of poor quality or whenever a minimum at a given (low) resolution occurs on a region of parameter space where there are no minima at the higher resolutions. It follows that it is possible for a test pattern to be correctly classified under the TD and erroneously classified under the MRTD. Quantifying how frequently such errors can occur is an important requisite for a complete understanding of the MRTD classifier.

For this, we start by noticing that the error rate achievable by any classifier is strongly dependent on how well the underlying representation separates the various image classes in the database. Ideally, all patterns from the same class should be confined to a region that does not overlap with the region containing the patterns from all other classes. Or, in other words, the distances between the patterns in the same class (to which we refer as *in-class* distances) should be small, while those between patterns in different classes (*out-of-class* distances) should be large. An interesting way to compare two classifiers is therefore to measure the ratios between their in-class and out-of-class distances.

Figure 8 presents a characterization of these ratios, for the TD and MRTD[2], under the conditions of experiment three and dataset D2. We measured all distances between test and training views and, for each test/train pair, the ratio D_{MRTD}/D_{TD}

[2]Qualitatively similar results were observed on identical experiments comparing the RMRTD and TD classifiers whose analysis is omitted for brevity.

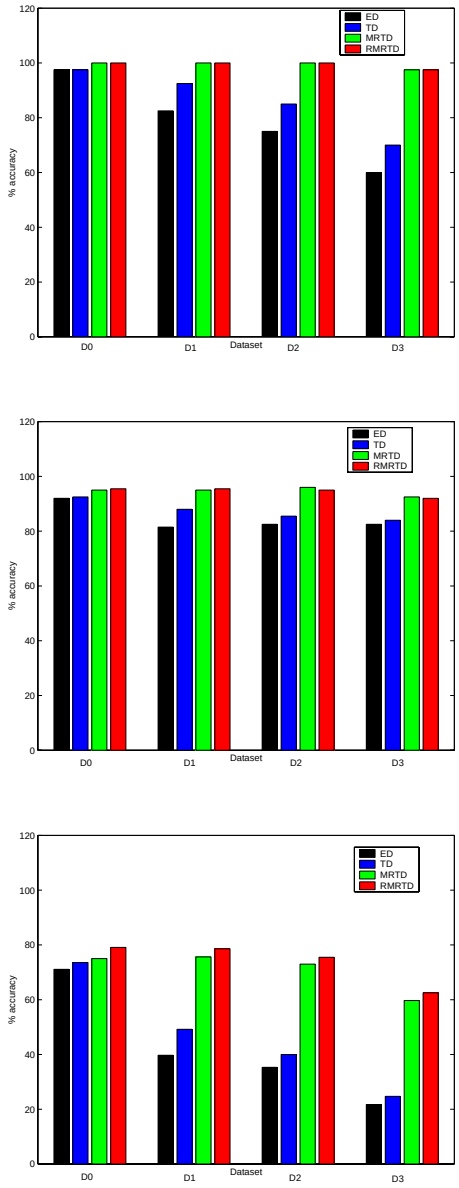

Fig. 7. Recognition accuracy. From top to bottom: results from the first, second, and third experiments. Datasets are ordered by degree of variability: D0 is the ORL database D3 is subject to the affine transformations of greater amplitude.

was computed. The figure presents the cumulative distribution function of this ratio for both the in-class and out-of-class distances, supporting two main conclusions. First, the MRTD is smaller than the TD with very high probability, in both in and out-of-class cases. However, the decrease is significantly more drastic in-class than out-of-class, e.g. 1) while 10% of the in-class ratios are smaller than 0.2 there are no out-of-class ratios of such magnitude, and 2) while the MRTD is smaller than half of the TD with probability 0.5 for in-class distances, the corresponding probability for out-of-class distances is only 0.1. Second, while the MRTD can lead to an increase of the in-class distances, the probability of such an event is very small (about 0.01). Together, these observations show that the MRTD separates the different classes significantly better than

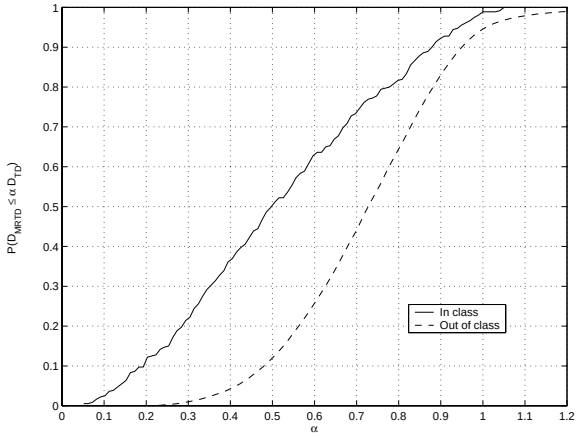

Fig. 8. Cumulative distribution function of the ratio D_{MRTD}/D_{TD} for both in- and out-of-class distances.

the TD.

Since better class separation usually translates into a smaller number of classification errors, the results of this experiment suggest that the probability of misclassification by the MRTD when the TD classifier is correct should be quite small. To confirm this conjecture, we collected statistics for the various types of MRTD/TD error combinations. These statistics, presented on table III, confirm that the percentage of MRTD errors which are not also TD errors is indeed quite small (5%). On the contrary, about 50% of the TD errors are not MRTD errors.

TD errors	228
TD errors which are not MRTD errors	101
MRTD errors	123
MRTD errors which are not TD errors	6

TABLE III
ERROR STATISTICS FOR THE TD AND MRTD CLASSIFIERS.

Figure 9 presents the 6 test patterns for which the TD classifier was correct and the MRTD counterpart in error. Analysis of the images reveals a common theme for these errors, which can be summarized by the following two conditions: 1) the mapping between the test pattern and the database pattern of the same class[3] cannot be captured by the affine model (e.g. due to out-of-plane rotation or changes in lighting) and 2) the (two-dimensional) pose of the database pattern of the correct class is quite close to that of the test pattern. While the first condition is responsible for the failure of the MRTD, the second enables the success of the TD. The fact is that whenever the transformation between the patterns deviates strongly from the affine model neither the TD or the MRTD should work. In these cases the reduced range of the TD prevents it from aligning out-of-class patterns that may belong or be very close to the manifold spanned by the test view but whose distance along the manifold is large. Consequently it will settle for patterns that are further away from the manifold, but whose manifold projection is close to the test pattern. While, typically, these matches are errors there is some (small) probability that they will belong to the right class. This, however, is the result of pure chance rather than any principled advantage of the TD classifier.

[3]Remember that in this experiment there is only one view of each face in the training set.

D. Robustness to deviations from the affine model

The artificially affine-transformed datasets of the previous section allow a precise quantification of the range of transformations over which the extensions of the TD hold. However, practical recognition usually involves image transformations that cannot be captured by a pure affine model, e.g. including some amount of illumination variation, out-of-plane rotation, or background clutter. For this reason, it is important to complement the evaluation above with a set of experiments performed on a database where the transformations are not imposed artificially. One example is the Media Laboratory's face database originally acquired to test the eigenfaces technique [30]. This database contains images of 16 subjects that vary in pose (head orientation), scale (camera zoom), and lighting, in a total of 27 images per subject (see Figure 10). It reflects a practical recognition scenario in the sense that no effort was made to keep the subjects from moving in between pictures, to precisely segment the faces from the background, or to precisely calibrate the variations in lighting, pose, etc.

To analyze how changes in the various imaging variables affect recognition accuracy we created two databases: a *neutral database* that contained one face from each subject in neutral position (upright face, head-on illumination, and medium scale) and a *non-neutral database* containing the remaining faces. We then considered each face in the non-neutral database as a query and ordered all faces in the neutral database according to their similarity to this query, measuring the average hit rate $\mathcal{H}_k$ for the top k matches, with k variable. The average hit rate is defined as

$$\mathcal{H}_k = \frac{1}{N}\sum_{i=1}^{N} 1\{\geq 1 \text{ hits in top } k \text{ matches of } i^{th} \text{ query}\}$$

where a hit is an image from the query subject and 1_x is 1 when x is true and 0 otherwise. This experiment simulates a common application scenario where a database is assembled under controlled conditions (e.g. a database of mugshots of convicted felons, or a photo gallery of a movie star) and is latter used for recognition in an uncontrolled scenario (e.g. airport surveillance or the detection of the scenes where the actor appears in a given movie).

Figure 11 a) presents the average hit rate for the ED, TD, MRTD, and RMRTD classifiers. It is clear that while the TD classifier is not much more accurate than the ED counterpart, a significant improvement can be achieved by using the MRTD or RMRTD. In fact the increase in hit rate of the two multi-resolution-resolution classifiers, relative to that based on the TD, is always larger than at least 10% (in absolute terms). When compared to the results from the previous sections, the only surprising aspect of the plots in the figure is the somewhat disappointing performance of the RMRTD, which slightly under-performs the MRTD. To explain this observation, as

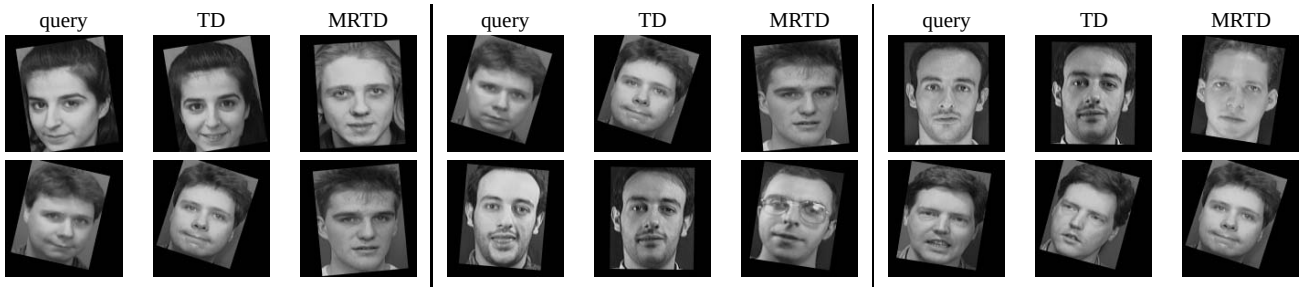

Fig. 9. Test patterns for which the TD classifier is correct and the MRTD classifier in error. Query is the text pattern, TD the best match under the TD, and MRTD the best match under the MRTD.

Fig. 10. The set of images from a subject in the MIT Media Laboratory's face database.

well as gaining some insight on the precise dependence of the MRTD on the imaging conditions, we performed a second set of experiments.

For this second set the goal was to analyze the impact of each type of transformation on the recognition accuracy. To fulfill this goal we considered each face in neutral database as a query and created five subsets of the non-neutral database, containing the following images:

- *rotation database* - all faces at the scale of the query, under the same illumination, but with different head orientation;
- *scale database* - all faces with the pose of the query, under the same illumination, but at different scale;
- *scale and rotation database* - all faces under the same illumination, but with different head orientation and scale;
- *illumination database* - all faces with the same head orientation and scale as the query, but different illumination;
- *scale, rotation, and illumination*- all faces with different head orientation, scale, and illumination than those of the query.

We then ordered all faces in each of these databases according to their similarity to each query. Figure 11 b)-f) presents plots of the resulting average hit rate for the ED, TD, MRTD, and RMRTD classifiers in each database.

Three interesting conclusions can be taken from the figure. The first is that both the MRTD and RMRTD perform better than the TD or ED in all but one case (the illumination database). In fact, as long as the illumination is the same for the query and database images the former outperform the latter by a significant amount. For example, the hit ratio for $k = 1$ (percent of the queries for which the first match is correct) of the MRTD or RMRTD is always more than twice that of the ED or TD. While undesirable, the degradation of recognition accuracy in the presence of illumination variability was expected, since illumination changes are beyond the scope of the affine model. Notice, nevertheless, that the MRTD and RMRTD do exhibit some robustness to these changes, for which retrieval accuracy is actually better than that achieved under rotation. What is surprising is the high robustness of the ED and TD under variable illumination conditions.

A second interesting aspect is the very different response of the MRTD and RMRTD to variations in scale and head orientation. While both perform very well under scaling transformations (the RMRTD actually achieves a perfect score of $\mathcal{H}_k = 1$ for all k considered) the recognition accuracy degrades considerably under rotation. This is, once again, a consequence of a significant deviation from the affine model. While under scaling both the face and the background are subject to the same affine transformation, for rotation only the subject's head is tilted (the background remains the same). Therefore, a single affine transformation cannot map the query into the database image. Nevertheless, while the performance of the MRTD and RMRTD degrades in the presence of head tilt, the gain over the TD or ED is still significant.

A third interesting observation is that while the performance of the RMRTD is equivalent to or better than that of the MRTD when the mapping between query and prototype is approximately affine, that is not the case in the presence of illumination variation. In fact, Figure 11 f) shows that the MRTD actually performs best when illumination variation, scaling, and rotation are all present. This confirms the results of Figure 11 a) and supports the following conclusion: while a robust estimator is advantageous when image transformations comply with the model used to the derive the TD (affine in

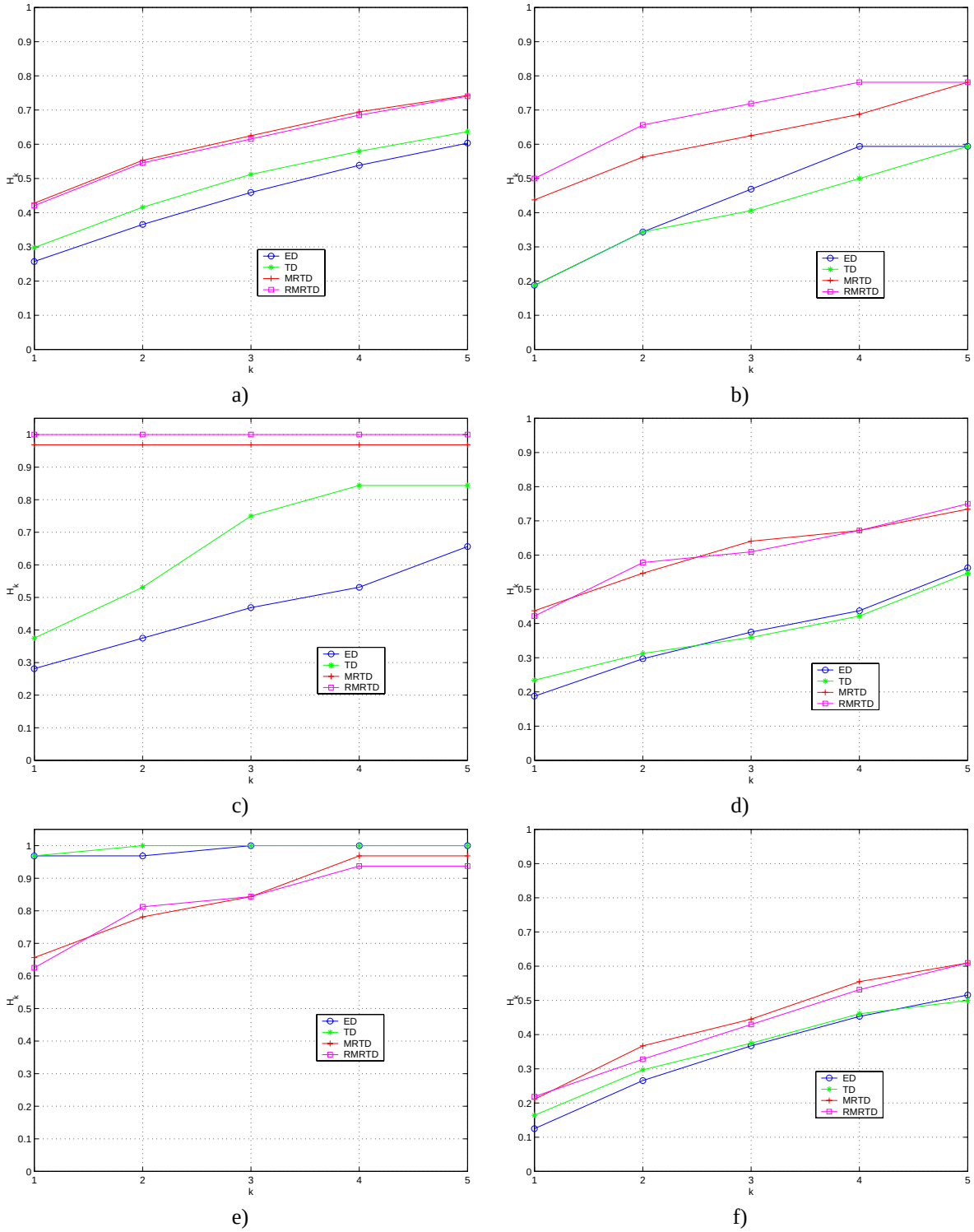

Fig. 11. Average hit rate as a function of the number of matches for the ED, TD, MRTD, and RMRTD classifiers on the various databases discussed in the text. While in a) the images in the non-neutral database were used as queries, in the remaining cases the queries came from the neutral database. Retrieved images came from: a) neutral, b) rotation, c) scaling, d) scale and rotation, e) illumination, and f) scale, rotation, and illumination database.

these experiments) it can, on the other hand, perform worse when this is not the case. Such a conclusion is consistent with what is known about robust estimators, namely that there is always a breakdown point in terms of the number of outliers above which they are likely to fail [9], [18].

Hence, while a robust classifier can be useful by, for example, ignoring background pixels when they do not conform with the affine model, it is important to guarantee that it will never operate above this breakdown point. In practice this can usually be achieved in two ways: 1) extending the transformation model on which the TD is based in order to cover all sources of variation present in the application of interest, or 2) pre-process all images to eliminate such variations. For example, illumination changes could be handled by 1) redefining the TD as

$$\mathcal{T}(M,N) = \min_{\mathbf{p},a,b} ||M(\mathbf{x}) - aT_{\mathbf{p}}[N(\mathbf{x})] + b||^2, \qquad (24)$$

where a and b are constants, or 2) using standard preprocessing tricks in common use in the face detection and recognition literature, such as subtracting a plane to the image intensities, performing histogram equalization, or cropping the images tightly around the face area [19], [27]. The latter solution would also eliminate the problems caused by the background, namely when it does not follow the transformation of the subject's face. The results obtained for the scaling database (where the affine model holds reasonably well for most of the image area considered in the matching) indicate that, if these steps are taken, a robust estimator should be sufficient to overcome the errors of the MRTD.

E. Implementation complexity

The previous sections illustrate a clear advantage, in terms of recognition accuracy, of the multi-resolution classifiers. However, this gain is achieved at the expense of increased computational complexity. The practical relevance of the multi-resolution classifiers can therefore only be assessed after an analysis of this computational penalty.

For this we notice that all the distances considered above involve, at some point, cycling through all the pixels in the query and database images. The operations carried inside this "pixel loop" include subtracting the two images (all that is required by the ED), collecting spatial derivatives, and computing running sums such as those of (17) or (19). Hence, the computational complexity of this pixel loop is $O(NM)$ where N and M are the image dimensions. The multi-resolution classifiers execute this loop inside a "multi-resolution loop", i.e. by repeating the estimation at each resolution level. If there are $L+1$ such levels and the images are sub-sampled by a factor of two (in each dimension) at each level, the overall complexity is

$$O\left(\sum_{i=0}^{L} \frac{NM}{4^i}\right) = O\left(\frac{4NM}{3}(1 - \frac{1}{4^{L+1}})\right),$$

a sequence that rapidly converges to $O(4NM/3)$. Hence, the computational increase of the multi-resolution classifiers is never larger than 33% of the computation required by the TD.

This, however, assumes that an exhaustive search is performed at each resolution. In practice it usually becomes clear, even at the lowest resolutions, that some of the images in the database will not be a good match to the query. These images can therefore be ignored in the subsequent levels of the multi-resolution decomposition without any degradation of recognition accuracy. Assuming that there are a total of S images in the database and, on average, only kS are retained at each level, $k \in (0,1]$, the overall search complexity will be

$$O\left(NMS\sum_{i=0}^{L}\frac{k^i}{4^{(L-i)}}\right) = O\left(\frac{NMS}{4^L}\frac{(4k)^{L+1}-1}{4k-1}\right).$$

The ratio of computation involved in the multi-resolution search over that required by the full resolution search is therefore $(4k)^{L+1} - 1/4^L(4k-1)$. Figure 12 shows the dependence of this ratio on k and L, making it clear that, for each L, there is a significant range of k for which the multi-resolution search is more effective than full search. For example when $L = 4$, the value which we have used in all experiments discussed above, this will hold as long as k is smaller than 0.92, i.e. even if only 8% of the images are discarded at each resolution.

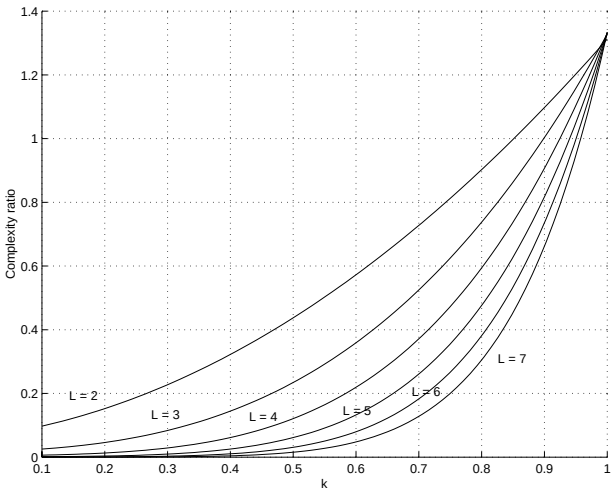

Fig. 12. Ratio of computation complexity between multi-resolution and single-resolution search as a function of the number of resolution levels (L) and the percentage of images retained at each level (k).

F. Semantic video classification

We finalize with some experimental evidence for the ability of the MRTD to capture elements of video structure that are important for its semantic classification. The goal is not so much to describe a full fledge semantic classifier or to show that one of the variations of the TD is much better than the others for this task, but to make clear that the MRTD is applicable in much broader settings than simple visual recognition. We emphasize that this is an important attribute when one considers the design of practical multimedia architectures, which must be versatile enough to be usable in diverse applications. For the MRTD, it stems from the fact that the MRTD has an objective, but intuitive, interpretation: it is a measure of the differences between two patterns that cannot

be canceled by alignment. Obviously, the meaning of these differences depends on the particular set of transformations that are allowed. In the affine case, they can be seen as the image differences that cannot be compensated by nulling out camera motions such as pans, zooms, or in-plane rotation.

This suggests that the MRTD could be used as a metric for the action in a scene. Once the camera motion is compensated, the differences that are left are likely to be due to the motion of objects in the scene. In general, the stronger the amplitude of the object motion and the larger the object size, the larger will be the TD. While the ability to detect action is an asset for various multimedia applications (from detecting events in interactive environments to retrieving the action scenes of a movie), it is hard to conceive that it could be done without at least compensating for camera motion. For example, pans are prevalent in scenic videos that depict scenes of very low activity. Hence, one could argue that the TD should be a dimension of any feature space used for detecting action. We next present evidence that, in fact, the TD by itself already appears to capture most of the information required for this detection. Since these results were already presented in [33], we will only summarize them here.

To evaluate semantic classification we relied on a database containing 23 promotional movie trailers for commercially released feature films. Each trailer consists of 2 to 5 minutes of video and the total number of shots in the database is 1959. The movie titles are presented in Table IV. Figure 13 shows how the movie database populates a feature space obtained by segmenting the video into shots and simply measuring the average duration of each shot and the average value of the MRTD (normalized to $[0, 1]$ by dividing by the maximum value along each axis) between consecutive frames in the shot. We also performed a search in the *Internet Movie Database* (IMDB) [1] for the *genre* assigned to each movie by the *Motion Picture Association of America*. Three major classes were identified: *romance/comedy*, *action*, and *other* (which includes *horror*, *drama*, and *adventure*). There were not enough points in the movie sample to further subdivide the other class in a meaningful way. The genre classes are indicated in the plots by the symbol used to represent each movie.

Several interesting observations can be made from the figure. First, the points seem to obey a law of the type $length \times activity = constant$. This is particularly interesting because the existence of a related law, $character \times action = constant$ has been postulated in the film theory literature [13]. This seems to confirm the fact that the MRTD is a good indicator for the *action* content of a movie. Second, there seems to be a clear separation between the three semantic classes in the activity/length feature space. In particular, movies of the *romance* and *comedy* genres are mostly above the top dashed line, *action* movies below the bottom one, and the other genres in between.

In fact, there are only four movies that violate these rules, "jungle", "madness", "blankman" and "edwood", and all correspond to cases where the semantic classification is ambiguous. For example, while the comedies above the top dashed line are typically categorized as *comedy/romance* or simply *comedy*, "edwood" receives the awkward categorization of *comedy/drama* (indicating that characterizing its content is probably a difficult task), and "blankman" that of *comedy/screwball/super hero* confirming the fact that it is an action-packed comedy, which could easily fall in the *action* category. Thus while, strictly speaking, the placement of these movies on the *other* and *action* classes is incorrect, it is semantically plausible. Similarly, while the romances above the top line either belong to the category *drama/romance* or *comedy/romance*, "jungle" is categorized as *adventure/romance* indicating a degree of action which is unusual for movies in the *romance* class. Finally, while "madness" is assigned to the *horror* genre, it is full of action-packed scenes. More samples from the *horror* class would be necessary for a deeper analysis of the interplay between these two genres.

We believe that these results illustrate how the natural decomposition into "what can be explained by camera motion" plus "what is left" is likely to play a role in semantic video analysis. A more thorough analysis of the semantic classification results is presented in [33], where we compared the MRTD to histogram intersection [28], the most commonly used similarity function in content-based image retrieval. It was shown that, although the classification rates were similar, histogram intersection led to to less intuitive errors. A good example is the movie "riverwilde", an action movie whose plot revolves around white-water rafting and contains numerous shots depicting this sport. While these shots exhibit a significant amount of motion from frame-to-frame, the color histograms tend not to change too much, because there is always plenty of water in the background. The action content cannot, therefore, be captured well by the histogram distance. On the other hand, since it cannot be explained by camera motion, it is captured by the MRTD.

TABLE IV
TITLES OF THE ENTRIES IN THE MOVIE DATABASE AND NAMES THAT APPEAR ON FIGURE 13.

Movie	Legend
"Circle of Friends"	circle
"French Kiss"	french
"Miami Rhapsody"	miami
"The Santa Clause"	santa
"Exit to Eden"	eden
"A Walk in the Clouds"	clouds
"While you Were Sleeping"	sleeping
"Bad Boys"	badboys
"Junior"	junior
"Crimson Tide"	tide
"The Scout"	scout
"The Walking Dead"	walking
"Ed Wood"	edwood
"The Jungle Book"	jungle
"Puppet Master"	puppet
"A Little Princess"	princess
"Judge Dredd"	dredd
"The River Wild"	riverwild
"Terminal Velocity"	terminal
"Blankman"	blankman
"In the Mouth of Madness"	madness
"Street Fighter"	fighter
"Die Hard: With a Vengeance"	vengeance

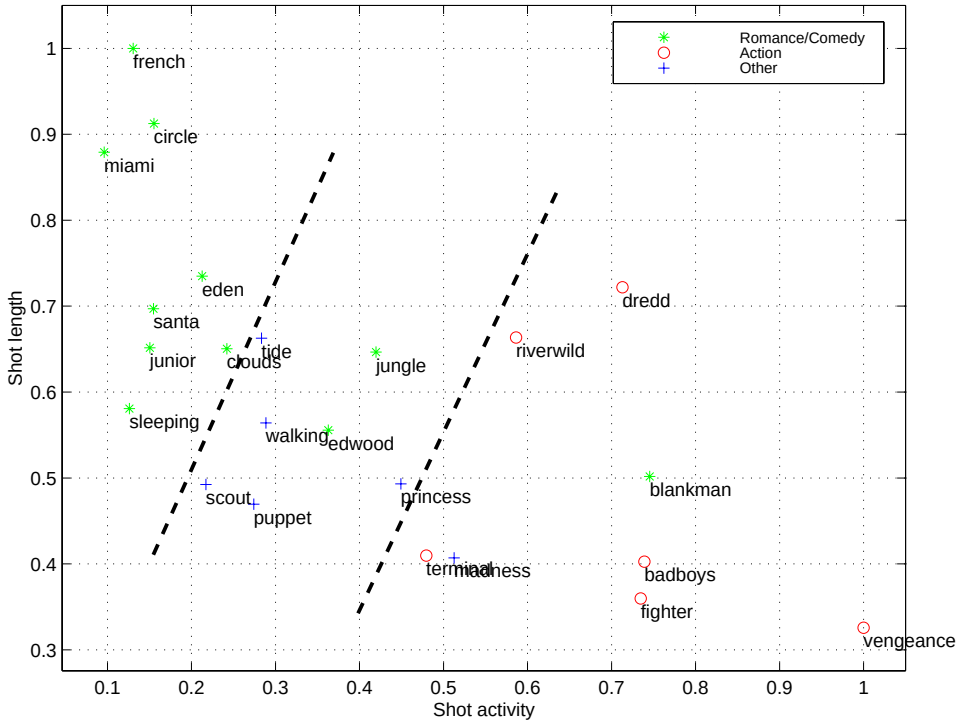

Fig. 13. Population of the feature space by the movies in our database. Movie names are listed in Table IV.

VII. Conclusions

In this work, we introduced the multi-resolution tangent distance. In the multimedia context, this distance has several interesting properties. First, it is generic and can be used for any task involving image similarity. For example, the MRTD classifier applies equal well to face recognition or detection, gesture and character recognition, recognition of traffic signals, video shot segmentation, etc. Despite its general purpose character, the MRTD achieves high classification rates, particularly for tasks where multiple views of each prototype pattern are available, and exhibits high invariance to linear transformations of the patterns to classify (that can impair significantly the performance of techniques based on the ED). It relies on a iconic (i.e. pixel-based) representation of the images to classify and does not, therefore, depend on features which are inherently task dependent, typically tricky to define, error prone, and many times expensive to compute and track. The MRTD can also be implemented with complexity equivalent to or smaller than that of the ED, is easily combined with robust estimation techniques, and is suited for hierarchical image analysis tasks.

In addition to recognition, the natural interpretation of the MRTD as what remains after camera motion is compensated makes it suited for various video analysis and classification tasks. We illustrated this fact with simple experiments on semantic movie classification, but the distance could also be applied to the segmentation of video into shots, the creation of image mosaics, or any application where the decomposition into camera and object motion is relevant. Despite all these good properties, the main advantage of the MRTD as a similarity metric may be of a practical nature. Because multimedia processors are required to support a wide array of applications, it is important that various tasks can share the same hardware architecture. The flexibility of the MRTD, shown here by its application to problems as diverse as face recognition and video classification, as well as the natural connection to motion estimation and mosaic creation can be significant assets in this context.

References

[1] *Internet Movie Database*. http://us.imdb.com/.

[2] P. Aigrain, H. Zhang, and D. Petkovic. Content-based Representation and Retrieval of Visual Media: A State-of-the-Art Review. *Multimedia Tools and Applications*, Vol. 3:179–202, 1996.

[3] P. Anandan, J. Bergen, K. Hanna, and R. Hingorani. Hierarchical Model-Based Motion Estimation. In M. Sezan and R. Lagendijk, editors, *Motion Analysis and Image Sequence Processing*, chapter 1. Kluwer Academic Press, 1993.

[4] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995.

[5] P. Burt and E. Adelson. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. on Communications*, Vol. 31:532–540, 1983.

[6] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[7] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[8] B. Frey and N. Jojic. Estimating Mixture Models of Images and Inferring Spatial Transformations Using the EM Algorithm. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado*, 1999.

[9] P. Huber. *Robust Statistics*. John Wiley, 1981.

[10] M. Irani, , P. Anandan, J. Bergen, R. Kumar, and S. Hsu. Mosaic Representations of Video Sequences and Their Applications. *Signal Processing: Image Communication*, 8(4), May 1996.

[11] B. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proc. DARPA Image Understanding Workshop*, 1981.

[12] J. Mao and A. Jain. Texture Classification and Segmentation Using Multiresolution Simultaneous Autoregressive Models. *Pattern Recognition*, 25(2):173–188, 1992.

[13] Wallace Martin. *Recent Theories of Narrative*, chapter 5. Cornell University Press, Ithaca, NY, USA, 1986.
[14] M. Massey and W. Bender. Salient Stills: Process and Practice. *IBM Systems Journal*, Vol. 35(3 and 4), 1996.
[15] D. Montegomery and E. Peck. *Introduction to Linear Regression Analysis*. John Wiley, 1992.
[16] H. Murase and S. Nayar. Visual Learning and Recognition of 3-D Objects from Appearence. *International Journal of Computer Vision*, 14:5–24, 1995.
[17] S. Ravela and R. Manmatha. Retrieving Images by Appearence. In *International Conference on Computer Vision*, 1998, Bombay, India.
[18] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. John Wiley, 1987.
[19] H. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.
[20] H. Rowley, S. Baluja, and T. Kanade. Rotation Invariant Neural Network-Based Face Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, California*, 1998.
[21] H. Sawhney and S. Ayer. Compact Representations of Videos Through Dominant and Multiple Motion Estimation. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, Vol. 18, August 1996.
[22] B. Schiele and J. Crowley. Recognition without Correspondence using Multidimensional Receptive Field Histograms. *International Journal of Computer Vision*, 36(1):31–50, January 2000.
[23] H. Schneiderman and T. Kanade. A Statistical Method for 3D Object Detection Applied to Faces and Cars. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, 2000.
[24] P. Simard, Y. Le Cun, and J. Denker. Efficient Pattern Recognition Using a New Transformation Distance. In *Proc. Neural Information Proc. Systems*, Denver, USA, 1994.
[25] P. Simard, Y. Le Cun, and J. Denker. Memory-based Character Recognition Using a Transformation Invariant Metric. In *Int. Conference on Pattern Recognition*, Jerusalem, Israel, 1994.
[26] S. Smoliar and H. Zhang. Video Indexing and Retrieval. In B. Furth, editor, *Multimedia Systems and Techniques*. KAP, 1996.
[27] K. Sung and T. Poggio. Example Based Learning for View-Based Human Face Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):39–51, January 1998.
[28] M. Swain and D. Ballard. Color Indexing. *International Journal of Computer Vision*, Vol. 7(1):11–32, 1991.
[29] Martin Szummer and Rosalind Picard. Indoor-Outdoor Image Classification. In *Workshop in Content-based Access to Image and Video Databases*, 1998, Bombay, India.
[30] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3, 1991.
[31] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image Classification for Content-Based Indexing. *IEEE Trans. on Image Processing*, 10(1):117–130, January 2001.
[32] N. Vasconcelos. *Bayesian Models for Visual Information Retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.
[33] N. Vasconcelos and A. Lippman. Statistical Models of Video Structure for Content Analysis and Characterization. *IEEE Trans. on Image Processing*, 9(1):3–19, January 2000.
[34] N. Vasconcelos and A. Lippman. A Bayesian Framework for Semantic Content Characterization. In *Proc. IEEE Computer Vision and Pattern Recognition Conf.*, Santa Barbara, California, 1998.
[35] J. Wang and E. Adelson. Representing Moving Images with Layers. *IEEE Trans. on Image Processing*, Vol. 3, September 1994.
[36] Y. Weiss. Smoothness in Layers: Motion Segmentation Using Nonparametric Mixture Estimation. In *Computer Vision and Pattern Recognition Conf.*, San Juan, Puerto Rico, 1997.
[37] H. Zhang, S. Smoliar, and J. Wu. Content-Based Video Browsing Tools. In A. Rodriguez and J. Maitan, editors, *Symposium on Electronic Imaging Science and Technology: Multimedia Computing and Networking*, pages 389–398, SPIE Vol. 2417, Feb. 1995, San Jose, California.

Nuno Vasconcelos received a *licenciatura* in electrical engineering and computer science from the Universidade do Porto, Portugal in 1988, a Master of Science from the Massachusetts Institute of Technology in 1993, and a PhD from the Massachusetts Institute of Technology in 2000. From 2000 to 2002, he was a member of the research staff at the Compaq Cambridge Research Laboratory, which in 2002 became the HP Cambridge Research Laboratory. In 2003 he joined the electrical and computer engineering department of the University of California, San Diego. He has worked in various areas including signal processing and compression, computer vision, machine learning and multimedia systems. His current interestest are in statistical signal processing, statistical computer vision, machine learning, large signal repositories, and multimedia.

Andrew Lippman received both his B.S. and M.S. degrees in electrical engineering from MIT. In 1995 he completed his Ph.D. studies at the EPFL, Lausanne, Switzerland. He is currently a Senior Research Scientist at MIT. He directs the MIT Media Lab's research consortium entitled "Digital Life" that addresses bits, people and community in a wired world. He holds eleven patents in television, digital image processing and interface technologies. His current research interests are in the design of systems for digital expression and entertainment and on global, interactive digital television infrastructures.