HAL
open science

Comparison of video dynamic contents without feature matching by using rank-tests

Alain Lehmann, Patrick Bouthemy, Jian-Feng Yao

To cite this version:

Alain Lehmann, Patrick Bouthemy, Jian-Feng Yao. Comparison of video dynamic contents without feature matching by using rank-tests. [Research Report] RR-5586, INRIA. 2005, pp.15. inria-00070421

HAL Id: inria-00070421

https://inria.hal.science/inria-00070421

Submitted on 19 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Comparison of video dynamic contents without feature matching by using rank-tests

Alain Lehmann — Patrick Bouthemy — Jian-Feng Yao

N° 5586

Mai 2005

Thème COG

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--5586--FR+ENG

Comparison of video dynamic contents without feature matching by using rank-tests

Alain Lehmann, Patrick Bouthemy, Jian-Feng Yao

Thème COG — Systèmes cognitifs
Projet VISTA

Rapport de recherche n° 5586 — Mai 2005 — 15 pages

Abstract: This report presents a novel and efficient dissimilarity measure between video segments. We consider local spatio-temporal descriptors. They are considered to be realizations of an unknown, but class-specific distribution. The similarity of two video segments is calculated by evaluating an appropriate statistic issued from a rank test. It does not require any matching of the local features between the two considered video segments, and can deal with a different number of computed local features in the two segments. Furthermore, our measure is self-normalized which allows for simple cue integration, and even on-line adapted class-dependent combination of the different descriptors. Satisfactory results have been obtained on real video sequences for two motion event recognition problems.

Key-words: dissimilarity measure, rank test, video sequence comparison

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex (France)
Téléphone : +33 2 99 84 71 00 —Télécopie : +33 2 99 84 71 71

Comparaison de contenus de vidéos par des tests de rang sans appariement de descripteurs

Résumé : Ce rapport présente une nouvelle mesure de similarité entre des séquences vidéo simple à évaluer. On considère des descripteurs spatio-temporels. Ils sont considérés comme des réalisations d'une distribution inconnue mais fonction de la classe de contenus considérée. La similarité de deux séquences vidéo est calculée par l'évaluation d'une statistique issue d'un test de rang. Il n'est pas nécessaire que les descripteurs locaux des deux séquences soient mis en correspondance ni que les nombres de descripteurs calculés dans les deux séquences soient égaux. De plus, notre mesure est normalisée automatiquement, ce qui permet d'utiliser des techniques simples pour l'intégration de différents descripteurs ainsi que des combinaisons de ces derniers fonction de la classe et adaptées en ligne. Des résultats satisfaisants ont été obtenus sur des séquences video réelles pour deux problèmes de reconnaissance d'événements liés au mouvement perçu.

Mots-clés : mesure de similarité, test de rang, comparaison de séquences vidéo

1 Introduction

Since the amount of multimedia data is rapidly growing, automatic systems are needed to process this huge amount of data. Therefore, the development of methods which are able to recognize semantically similar things is crucial to handle these data. Such methods must be applicable in the context of video databases to group similar video segments together, to satisfy queries, to extract highlights (e.g., in sport videos), or to summarize videos, but also in the context of video surveillance to detect special types of events which probably require human intervention. Hence, the challenging problem to solve is to map semantical concepts and low-level video features due to the well-known semantic gap [1, 7, 6].

To perform a recognition task, one has to define the representation of the video segment, which is of crucial importance. Two complementary approaches can be differentiated. There are the ones which extract global features from the whole video segment. They are simple to implement, but may have problems with complex scenes. An example for motion recognition using a global approach is [10] where a simple non-parametric distance measure based on histograms of spatial and temporal intensity gradients has been used. In [8], sport videos have been characterized with probabilistic motion models to capture the dominant image motion (i.e. the camera motion) and the residual scene motion. These models are learnt from global occurrence statistics computed over the whole video segment. Maximum likelihood criterion were used within a supervised classification scheme.

On the other side, there are the local approaches which extract features from spatio-temporally localized regions to overcome the problems of the global approach. A difficulty of these local approaches is, however, that the segments are no longer represented by a single feature vector, but a set of feature vectors. As a consequence, the comparison of segments is no longer straightforward and is normally achieved by matching the local features between the processed video and the database. In [5], a set of local space-time descriptors for recognizing motion patterns are presented and evaluated. The matching of the features is done in a greedy manner, whereas different distance measures have been tested. Videos of human motion activities have been used for the tests and good recognition rates have been reported.

In our method, we have considered two types of motion descriptors for characterizing dynamic content of a video sequence with local features. The first one accounts for the spatio-temporal evolution of the interest points which is assumed to be related to the trajectory of the moving objects. The second tries to capture the intensity of the motion, and we actually adapted the scene motion model of [8] to our local setup. We have designed an original dissimilarity measure which does not need an explicit pairing of the local feature vectors of the two segments and can deal with a different number of computed feature vectors in the two segments to be compared. It is based on a simple statistical test and is easy to compute as it involves ranking operations only. The assumption is, that a feature value set is generated by an unknown stochastic process. The basic idea is then to test whether the two feature value sets are generated from the same process or not. Furthermore, we can combine the different descriptors in a class-dependent adaptive way.

As the description of motion events is a non-trivial problem, a single descriptor is indeed not sufficient and a set of descriptors has to be used. Boosting [2] has become a popular method for automatic feature selection or combination. However, the generalization to multi-class classification is not obvious, even if some investigations have been undertaken. [3, 9]. A disadvantage of these boosting algorithms is however their computationally expensive learning stage. We have defined a simple method which is able to learn the quality of the individual descriptors to discriminate a given class from the remaining ones. This descriptor quality is then exploited to deduce a proper weighting to combine the different descriptors in the designed statistical test.

The remainder of this report is organized as follows. The classification framework comprising the statistical dissimilarity measure and the cue integration technique is introduced in Section 2. In Section 3, the interest point selection stage and the considered local motion features are presented. Finally, experimental results of motion event classification are reported in Section 4. Concluding remarks are given in Section 5.

2 Classification Framework

2.1 Dissimilarity measure between two video segments

The task of motion event recognition can be seen as the problem of classifying a given video segment according to some predefined classes $c \in \mathcal{C}$. To achieve this task we are previously given several examples s for each class, for which we know the class membership, i.e. $C(s) \in \mathcal{C}$. This set of examples will be further denoted as video database $\mathcal{S}$. The problem of event recognition can be formulated as the search for the minimum of a dissimilarity function Φ:

$$\hat{C}(r) = \arg\min_{c \in \mathcal{C}} \Phi(r|c) \tag{1}$$

where r is the tested video segment. We now reformulate the classification problem as a retrieval problem, i.e, we try to find the most similar segment s^* in our database S and base the classification on the class of s^*.

$$\hat{C}(r) = C(s^*), \quad \text{where } s^* = \arg\min_{s \in \mathcal{S}} \Phi(r, s|C(s)) \tag{2}$$

Actually, we consider the three best segments using majority voting to increase the robustness. In case that all three segments belong to different classes, the class of the best segment is chosen. In order to find the most similar segment we have to define a dissimilarity measure for two given segments r and s which may be class dependent. This class dependency can be justified by the fact that not all given descriptors have to be characteristic for all given classes. However, we have to ensure that the different dissimilarity measures are comparable in terms of their magnitude such that a segment which is more similar than one of another class also gets a smaller value.

Before we can define a dissimilarity measure, we have to specify how we represent a given video segment. As stated in the introduction, we use local features to characterize the segment content, that is a set of feature values. The considered local spatio-temporal descriptors will be introduced in Section 3 along with the technique to select the spatio-temporally localized regions of interest where these local descriptors are computed.

As one single descriptor is not sufficient to capture the complex notion of a motion event, we indeed consider a set of $d = 1 \dots L$ different local descriptors. Hence, we also have to specify a way to combine the dissimilarity values of the different descriptors. We consider a weighted sum, and the dissimilarity measure between two video segments r and s is finally given by

$$\Phi(r, s|(C(s)) = \sum_{d=1}^{L} \omega_d(C(s)) T_d(r, s) \tag{3}$$

where T_d is a similarity test which is now defined in the subsequent subsection 2.2 and ω is a class-dependent weighting which will be explained in subsection 2.3.

2.2 Wilcoxon Rank-Sum Test

In a method using local descriptors, a video segment is represented by a *set* of m feature vectors of dimension L (the number of considered descriptors), where m is the number of selected interest points in the video segment. With other words, we are confronted with the problem of comparing two sets of feature values Θ_r, Θ_s of not necessarily equal size m_r, m_s.

In contrast to comparing feature vectors, the comparison of sets of feature vectors is not obvious. A possible way is to establish correspondence between the features of the two sets by pairing the ones which are nearest in feature space (see Fig. 1(a)). However, there are

(a) No multiple assignemnts (b) Degenerated mapping

Figure 1: Illustration of possible mappings between feature sets and their problems

several problems using such a mapping. If we restrict the mapping such that each feature is used only once then there are some features which have not been considered (as the two sets do not have the same size) and hence we loose some of the extracted information. However, if we allow multiple assignments the problem of a degenerated mapping may arise where all features of one set are associated with one single feature of the other set (see Fig. 1(b)). The consequence is again that we do not consider all features. A further possibility may be to constrain the mapping somehow such that all features are equally considered, but it is

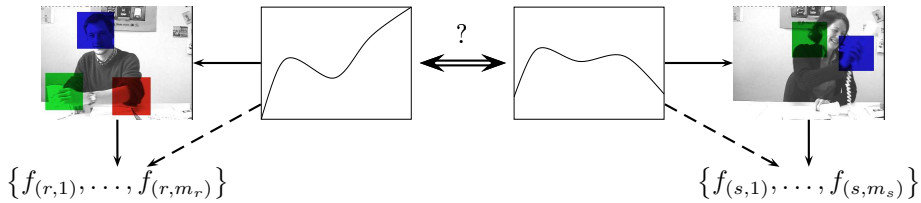

Figure 2: The video sequences are considered to be the result of an unknown probabilistic process. Consequently, we assume that the extracted features are realizations of an unknown, but class-dependent distribution.

not obvious how to do that. Furthermore, we would have to calculate the distance between every possible pair of features to find the nearest feature in the other set. This is quite expensive and hence a further drawback of such an approach.

Instead of establishing correspondence between the elements of the two sets, we consider the values as realizations of a unknown distribution $\mathcal{D}$ (see Fig. 2). Hence, to decide whether two segments r,s are of the same class it is sufficient to decide, whether the values are drawn from the same unknown distribution or not, i.e. testing the hypothesis $\mathcal{H}_0 : \mathcal{D}_r \equiv \mathcal{D}_s$. Let us first consider the case of a single local descriptor ($L = 1$).

The two-samples Wilcoxon Rank-Sum Test is a well-known[4] and statistical method to test this hypothesis $\mathcal{H}_0$ for scalar values. This non-parametric test has the advantage to be distribution-free and avoids the fit of a precise model. Specifically, the test statistic is given, in a normalized form, by

$$W_d = \left(\frac{12(m_r + m_s)}{m_r m_s}\right)^{\frac{1}{2}} \sum_{j=1}^{m_r} \left(\frac{R_{(m_r+m_s)j}}{m_r + m_s + 1} - \frac{1}{2}\right) \tag{4}$$

where $R_{(m_r+m_s)j}$ is the rank of the j-th value of the first feature set in the combined feature set, i.e. the position in the ordered sequence of both value sets. The distribution of this test statistics will converge for $m_r, m_s \to \infty$ to the $\mathcal{N}(0,1)$-distribution if the hypothesis $\mathcal{H}_0$ is fulfilled. If not, the value will be far from zero.

Hence, we define $T_d = W_d^2$ as an indicator of dissimilarity. As a consequence of the convergence result (if $\mathcal{H}_0$ is fulfilled), T_d is independent of the magnitude of the compared feature values. This property is important when we consider a class-dependent weighting as it automatically ensures the normalization of the dissimilarity value for all considered descriptors.

2.3 Feature combination by weighting

As the Wilcoxon Rank-Sum Test is only defined for scalar data, we cannot apply the test directly on the descriptor tuple. Instead, we treat each local descriptor separately as an individual descriptor and we will then introduce a combined dissimilarity measure.

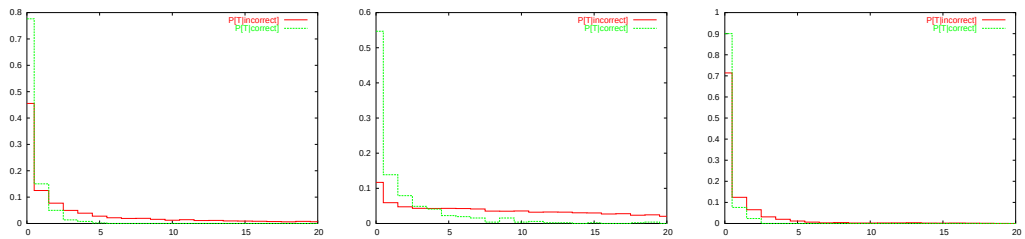

Figure 3: Examples for the conditional probabilities $p_d(t|c)$ (green line) and $p_d(t|\bar{c})$ (red line) obtained from the datasets. All of them show a pronounced peak at zero and a decreasing nature for increasing values.

The simplest way to combine the dissimilarity of all individual descriptors would be to just sum them, i.e., a uniform weighting. However, not all descriptors have to be characteristic for all classes and hence, a uniform weighting may arbitrarily degrade the quality of the dissimilarity measure. Instead, we try to learn the discriminative power of each individual descriptor d from the training data and deduce a proper weighting function ω for our dissimilarity measure.

An indicator for the appropriateness of a descriptor d is the success probability $\mathcal{P}\left[t_c < t_f|d\right]$ that the test result t_c for a segment pair yielding a correct classification is smaller than the value t_f for a segment pair yielding a false one. We first introduce the conditional probability density of the test statistics T_d given the classes of the tested segment pair. More precisely, we consider the two cases where either both segments are of the same class c or that one is of class c and the other one is *not* of class c, i.e.

$$\begin{aligned} p_d(t|c) &= \mathcal{P}[T_d(r,s) = t|C(r) = C(s) = c] \\ p_d(t|\bar{c}) &= \mathcal{P}[T_d(r,s) = t|C(r) = c, C(s) \neq c] \end{aligned} \tag{5}$$

Hence, the success probability of a descriptor d given a class c can be calculated as

$$\begin{aligned} \mathrm{Succ}(d|c) &= \mathcal{P}(t_c < t_f | t_c \sim p_d(\cdot|c), t_f \sim p_d(\cdot|\bar{c})) \\ &= \int_0^\infty p_d(t_c|c) \int_{t_c}^\infty p_d(t_f|\bar{c}) dt_f dt_c \end{aligned} \tag{6}$$

The analysis of the empirical histograms of the two conditional distributions (obtained from the training set) has shown that there is a pronounced peak at zero (see Fig. 3). Inspired by [8], we estimate these distributions using a mixture of a Dirac pulse and an exponential distribution

$$\widehat{p}_d(t|c) = \lambda_d(c)\delta_0(t) + (1 - \lambda_d(c)) \frac{1}{\beta_d(c)} e^{\frac{-t}{\beta_d(c)}} \mathbb{1}_{(t>0)} \tag{7}$$

which introduces some implicit smoothing which prevents overfitting to the data. The estimate for λ is defined as the fraction of samples $t < 1$. The estimate for β is simply the sample mean of all values $t \geq 1$. The formula for $\hat{p}_d(t|\bar{c})$ is analogue. The equation for the success probability can now be calculated as:

$$\mathrm{Succ}(d|c) = \lambda_d(c) + (1 - \lambda_d(c))(1 - \lambda_d(\bar{c}))\frac{\beta_d(\bar{c})}{\beta_d(c) + \beta_d(\bar{c})} \tag{8}$$

To keep the final dissimilarity measures comparable across different classes, we have to normalize these success probabilities, so that the sum over all descriptors is one. However, as the differences of these probabilities are rather small, we apply a exponential stretchting prior to the normalization, i.e.,

$$\omega_d(c) = \frac{1}{Z} e^{\alpha \mathrm{Succ}(d|c)} \qquad \text{with } Z = \sum_d e^{\alpha \mathrm{Succ}(d|c)} \tag{9}$$

where we empirically found $\alpha = 10$ to sufficiently stretch the rather small differences of these probabilities.

3 Local Motion Features

3.1 Interest Point Selection

The interest selection method which we have used in our experiments is based on a simple criterion to spot out the regions with a high density of scene motion activity, i.e.,

$$A(p,t) = \frac{1}{|\mathcal{F}|} \sharp \left\{ (q,\tau) \in \mathcal{F}(p,t) \quad | \quad |FD(q,\tau)| > \gamma \right\} \tag{10}$$

where $\mathcal{F}$ is a spatio-temporal neighbourhood window (e.g. $15 \times 15 \times 3$) and the FD is the temporal frame difference. We can also accomodate the case of a moving camera by first compensating the dominant image motion (e.g., represented by an affine motion model) which can be usually assumed to be related to the camera motion. Then, it can be cancelled by considering the DFD values (Displaced Frame Difference) instead of the FD values. Concerning the threshold, we used $\gamma = 80$ which results in a highly selective process.

The actual selection of the points is achieved in a greedy fashion, where we successively select the point with maximal criterion value. To avoid that all points are selected from about the same positions, and hence, to ensure that they are sufficiently spread over the whole video segment, we successively mask the surrounding of every selected interest point. Accordingly, the criterion $A(p,t)$ is explicitly modified after each selection in the sense that all FD values in a neighbourhood $\mathcal{F}'$ of the selected interest point are set to zero. In our experiments, this mask $\mathcal{F}'$ has been chosen such that the blocks $\mathcal{B}$ which are introduced in subsection 3.3 do not overlap.

The selection process is stopped as soon as the criterion value falls below a threshold relative to the initial criterion maximum, i.e., $\beta \max_{(p,t)} A(p,t)$, while we used $\beta = 0.01$ in the experiments.

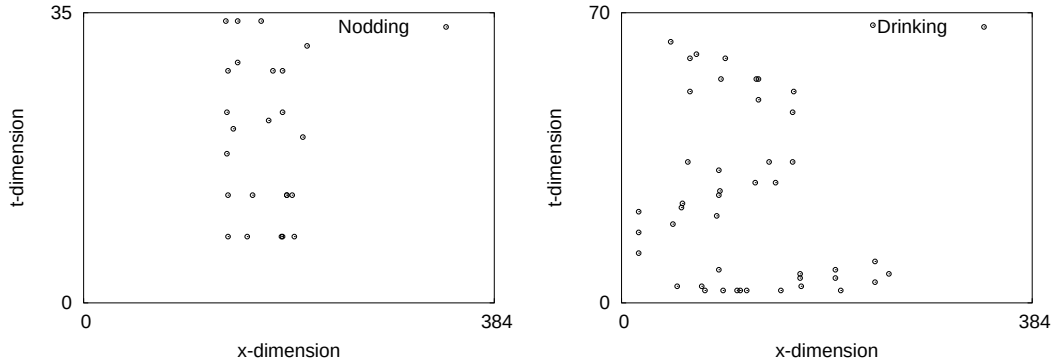

Figure 4: Illustration of the point clouds (projection onto the xt-plane) for the classes "nodding" (left) and "drinking" (right). In case of "drinking", the subject grabs the glass of water, drinks and puts it back which results in the shown left-right-left pattern. As "nodding" is just a rotation of the head, such a pattern cannot be observed.

3.2 Trajectorial Information

To characterize motion content in video, it seems naturally to somehow describe the trajectory of the moving objects in the image sequence. However, the detection of objects and estimation of their trajectories is a non trivial problem. Hence, we use a much simpler descriptor which tries to capture the spatio-temporal evolution of the interest points. We consider the 3D point cloud (see Fig. 4) generated by the selected interest points in the volume formed by the image sequence. Since we are not interested in the absolute position of the moving objects (as we do not assume the video segments to be aligned), but only at their spatio-temporal evolution, we will refer the measurements relativ to the center of gravity of these points.

To describe the point cloud, we consider for each interest point $p_i = (x_i, y_i, t_i)$ the following measurement

$$\nu(p_i) = (x_i - \bar{x})^g (y_i - \bar{y})^h (t_i - \bar{t})^l \tag{11}$$

of order $o = g + h + l$, where $(\bar{x}, \bar{y}, \bar{t})$ is the center of gravity of all interest points. Hence, we can calculate a feature vector for each interest point p_i where the different components or descriptors correspond to different combinations of g, h and l. In our experiments, we consider combinations with $o = 1 \ldots 6$.

Even though these measurements were inspired from the calculation of moments, there are some subtle differences. If we would compare video segments using moments, i.e. the sum of the computed values, we would run into the problem that the variation of the differences of moments of different orders are not equal, which makes the definition of an appropriate distance measure problematic. This however is crucial for the class dependent weighting. A

second difference is that the first order barycentric moments are always zero by definition and do not yield any information. In contrast, the first order measurement we consider have shown up to be discriminative (by looking at the weightings).

3.3 Motion Intensity Information

The motion intensity (velocity) is another important source of information to characterize motion events. If we consider for example a walking and a running person, the trajectory descriptor could probably not be very different, whereas a velocity-related descriptor should be. We have adapted the scene motion characterization introduced by [8] to our local approach. The histogram of the considered low-level motion features is no longer computed over the whole video segment, but in a block $\mathcal{B}_i$ of size $32 \times 32 \times 5$ surrounding the interest point p_i, where the considered motion feature is the averaged normal flow magnitude

$$\bar{v}(p,t) = \frac{\sum_{q \in \mathcal{W}(p)} \|\nabla I(q,t)\|^2 \cdot |v_n(q,t)|}{\max\left(|\mathcal{W}|\eta^2, \sum_{q \in \mathcal{W}(p)} \|\nabla I(q,t)\|^2\right)} \tag{12}$$

where $\mathcal{W}$ is a 3×3 neighbourhood window, η^2 is a noise related threshold and $v_n = \frac{-\partial I/\partial t}{||\nabla I||}$. Again, we could accomodate camera motion, if any, by considering the residual normal flow magnitude.

As proposed in [8], this distribution is modeled with a mixture distribution of a Dirac pulse at zero (corresponding to the symbolic state "no motion" and a continuous part representing the real motion values. In contrast to [8], the continuous part is modeled with a log-normal distribution, because there where observations where the zero-centered Gaussian restricted to $(0, \infty)$ was no longer suitable (see Fig. 5). It can be explained by the fact that the blocks are placed on regions with rather high motion activity. We get

$$\mathcal{P}[v|\mathcal{B}_i] = \lambda\delta_0(v) + \frac{1-\lambda}{v\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\log \frac{v}{m}}{\sigma}\right)^2} \mathbb{1}_{(v>0)} \tag{13}$$

The maximum likelihood estimation of the parameters is

$$\hat{\lambda} = \frac{1}{M} \sum_{v \in \mathcal{B}} \mathbb{1}_{(v \leq \varepsilon)} \tag{14}$$

$$\hat{m} = exp(\hat{\mu}) \qquad \text{with } \hat{\mu} = \frac{1}{\bar{M}} \sum_{v \in \mathcal{B}, v > \varepsilon} \log v \tag{15}$$

$$\hat{\sigma}^2 = \left(\frac{1}{\bar{M}} \sum_{v \in \mathcal{B}, v > \varepsilon} \log^2 v\right) - \hat{\mu}^2 \tag{16}$$

where M is the total number of samples in the block $\mathcal{B}_i$ and $\bar{M}$ is the number of samples with $v > \varepsilon = 0.1$.

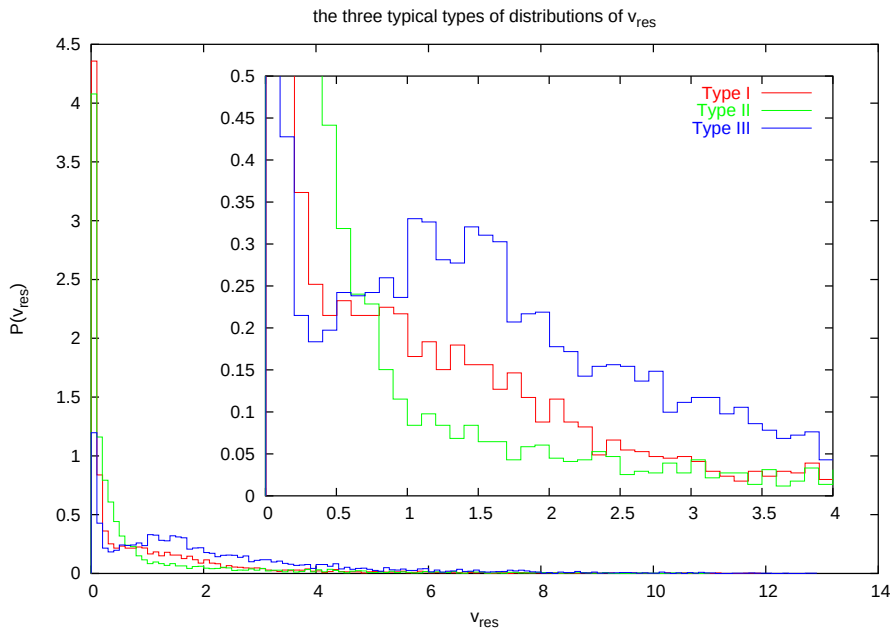

Figure 5: Some typical examples of occurrence distributions $\mathcal{P}[\bar{v}|\mathcal{B}]$ obtained from the tested video databases.

4 Motion Event Classification

The performance of our method has been evaluated on two different video databases for two event recognition problems. As the camera is fixed in both cases, we do not have to compensate for camera motion. For the evaluation we used a leave-one-out validation strategy: Each video segment has been classified based on the remaining ones. The set of local descriptors includes for each interest point (or block) the trajectory descriptors up to order 6 and the three motion intensity descriptors λ, m, σ (or part of them according to the experiments carried out).

4.1 Gesture Video Database

The first considered database consists of human gestures (see Fig. 6). There are six different classes, i.e., "shaking one's head", "nodding", "clapping hands", "answering the phone" and "drinking water". All these gestures have been carried out several times by seven different subjects. The total size of the database is 211 video segments. Furthermore, it has to be noted that videos are all recorded from the same viewpoint and the subject are centered in the screen. As all the gesture of the same class of an individual subject resembled each

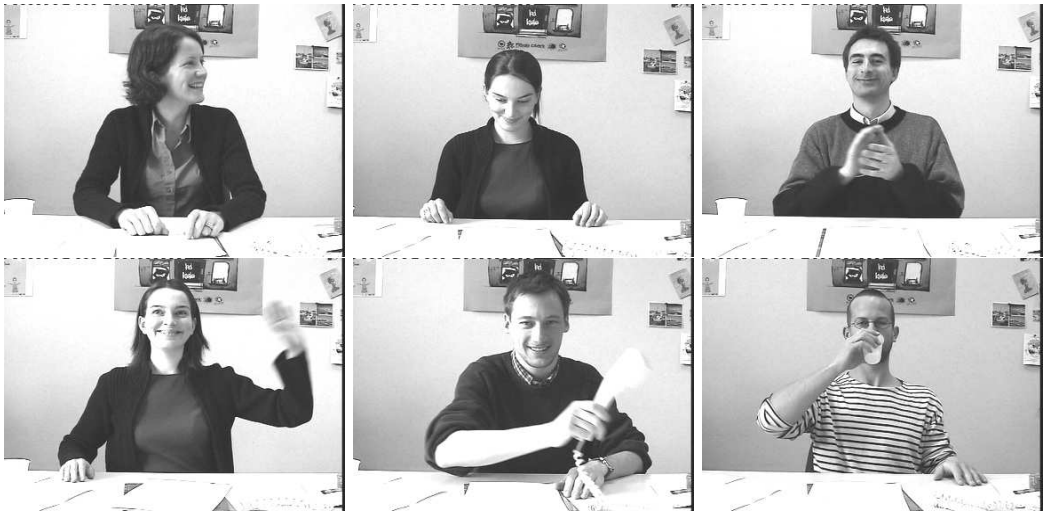

Figure 6: Example for each class of the "Gesture" database

	Shaking	Nodding	Clapping	Waving	Phoning	Drinking
Shaking	**87.5**	6.2		6.2		
Nodding	16.1	**64.5**		6.5	3.2	9.7
Clapping	3.2	3.2	**77.4**	16.1		
Waving	1.8	3.5	10.5	**82.5**		1.8
Phoning		6.5		6.5	**87.1**	
Drinking				3.4	6.9	**89.7**

average: 81%

	Shaking	Nodding	Clapping	Waving	Phoning	Drinking
Shaking	**84.4**	12.5		3.1		
Nodding	3.2	**74.2**	6.5	12.9		3.2
Clapping			**96.8**	3.2		
Waving		1.8	1.8	**96.5**		
Phoning	3.2		6.5		**90.3**	
Drinking					6.9	**93.1**

average: 89%

Figure 7: Confusion matrices for the "Gesture" video database with (left) uniform and (right) class-dependent weighting.

other very much, the validation has further been constrained, so that we not only exclude the current test segment, but all video segments of the same subject.

First, we tested our method using only the trajectory descriptors. The obtained results are reported in Figure 7. To show the influence of our class-dependent weighting scheme, the results using just uniform weighting are shown in the left column. It can be seen, that there is a rather large confusion between the classes "shaking" and "nodding" and also between "clapping" and "waving hands". The first confusion is rather evident as the moving object in both gestures, i.e. the head, stays at the same position and hence, the point cloud is

rather compact and shows no significant spatio-temporal evolution. The later one may be explained by the fact that the main motion of both gestures is horizontal. Furthermore, our simple descriptor cannot reveal whether there is one or several moving objects, i.e. both hands in case of "clapping" or one hand in case of "waving".

The results using class-dependent weighting are shown in the right column of Figure 7. As for the classes "clapping" and "waving hands", the confusion has been mostly resolved. Due to the similarity of the two first gesture types (explained before) there still remains a slight confusion between them. Looking at the overall performance, this class-dependent weighting significantly increased the classification rate by 8% to 89%.

The results using the descriptors capturing information about the motion intensity of the movement are left out as they did not yield any significative further improvements. The fact that all gestures are carried out at more or less the same velocity may explain why these additional descriptors are not capable to increase the performance in this experiment.

4.2 Basketball Video Database

The "Basketball" database (see Fig. 8) consists of 228 video segments where three classes, i.e. "shot on the basket", "lay-up" and "one-on-one" are differentiated. In contrast to the "Gesture" videos, the videos are taken from a variety of view points, and hence, there is no alignment anymore. However, the camera is still not mobile. The difficulty of these videos is that the intra-class variability is rather high as the movements (especially in case of "one-against-one") are not as clearly defined as for example for "clapping hands".

Figure 8: Example for each of the three classes of the "Basketball" database: Shot on the basket (left), lay-up (middle) and one-on-one (right).

Again, we considered the classification performance with uniform and class-dependent weighting (see Fig. 9) to show the influence of the later one. We get a the rather poor classification rate in case of class "one-on-one" using the uniform weighting. It may be due to the fact that there is always a shot on the basket at the end of the "one-on-one" video segments. As in the first test, the class-dependent weighting is able to correct a lot of misclassifications (while introducing just a very few ones). The overall classification performance is again increased by about 8%.

	Shot	Layup	1–1
Shot	**99.1**	0.9	0.0
Layup	3.1	**89.2**	7.7
1–1	34.7	20.4	**44.9**

average: 78%

	Shot	Layup	1–1
Shot	**98.2**	0.9	0.9
Layup	1.5	**89.2**	9.2
1–1	16.3	12.2	**71.4**

average: 86%

	Shot	Layup	1–1
Shot	**99.1**	0.0	0.9
Layup	0.0	**95.4**	4.6
1–1	16.3	8.2	**75.5**

average: 90%

Figure 9: Confusion matrices for the "Basketball" video database. Uniform weighting (left), class-dependent weighting (middle) and class-dependent weighting with all descriptors (right).

In contrast to the gesture sequences, the basketball video classes involve a large variability in terms of motion intensity (e.g., sudden movement in case of a dribbling) which can be exploited by the descriptors which characterize the intensity of the movement. The left column in Fig. 9 contains the results which we obtained including the motion intensity descriptors. As expected, the additional descriptors are capable to improve the classification rate further to 90%.

5 Summary and Conclusions

We have proposed a novel dissimilarity measure between video segments for local descriptors based on the Wilcoxon Rank-Sum test. This measure can be computed very efficiently, and it does not require any pairing of the features of the compared video segments an can straightforwardly handle a different number of feature values (i.e., interest points) per segment. Another property of this measure is that in case of similarity, the distribution of the measure converges to $\mathcal{N}(0,1)$ which can be seen as a self-normalization. This simplifies the integration of several descriptors. Furthermore, we have defined a way for learning the discriminative power of the different descriptors and for deducing a self-adaptive combination of the descriptors which performs significantly better than a uniform combination. The proposed framework has been tested on two motion classification problems where quite satisfactory results have been obtained using simple local motion features related to object trajectory and scene motion intensity observed in the image sequence. The proposed video segment similarity criterion can in fact be applied to any kind of features for video comparison, video classification or video retrieval.

References

[1] A.A. Efros, A.C. Berg, G. Mori and J. Malik Recognizing action at a distance. In *IEEE Int. Conf. on Computer Vision*, Nice, France October 2003

[2] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[3] V. Guruswami and A. Sahai. Multiclass learning, boosting, and error-correcting codes. In *COLT '99: Proc. of the twelfth annual Conf. on Computational Learning Theory*, pages 145–155. ACM Press, 1999.

[4] J. Hájek and Z. Šidák. *Theory of rank tests.* Academic Press, New York, 1967.

[5] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *SCVMA'04: Int. Workshop on Spatial Coherence for Visual Motion Analysis*, Prague, May 2004.

[6] Y.-F. Ma and H.-J. Zhang. Motion pattern-based video classification retrieval. EURASIP Journal on Applied Signal Processing, 2:199–208, 2003.

[7] O. Masoud and N. Papanikolopoulos. A method for human action recognition. *Image and Vision Computing Journal*, 21:729–743, 2003.

[8] G. Piriou, P. Bouthemy, and J-F. Yao. Extraction of semantic dynamic content from videos with probabilistic motion models. In *European Conf. on Computer Vision, ECCV'04*, Prague, May 2004 , Vol. LNCS 3023, Springer.

[9] R.E. Schapire. Using output codes to boost multiclass learning problems. In *ICML '97: Proc. of the Int. Conf. on Machine Learning*, 1997.

[10] L. Zelnik-Manor and M. Irani. Event-based video analysis. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii, volume 2, pages 123–130, December 2001.

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399