A Rejection-Based Method for Event Detection in Video

Margarita Osadchy and Daniel Keren

***Abstract*—This paper offers a natural extension of the newly introduced "anti-face" method to event detection, in both the gray-level and feature domains. For the gray-level domain, spatio-temporal templates are created by stacking the individual frames of the video sequence, and the detection is performed on these templates. In order to recognize the motion of features in a video sequence, the spatial locations of the features are modulated in time, thus creating a one-dimensional vector which represents the event in the detection process. The following applications are presented: 1) detection of an object under three-dimensional (3-D) rotations in a video sequence simulated from the COIL database; 2) visual recognition of spoken words; and 3) recognition of two-dimensional and 3-D sketched curves. The technique is capable of detecting 3-D curves in viewing directions which substantially differ from those in the training set. The resulting detection algorithm is very fast and can successfully detect events even in very low resolution. Also, it is capable of discriminating the desired event from arbitrary events, and not only from those in a negative training set. Possible applications of the techniques offered in this paper are in man–machine interaction, surveillance, and search and summarization in video databases.**

***Index Terms*—Event detection, rejectors, visual speech recognition.**

I. Introduction

WE PRESENT in this paper a new approach to event detection which is fast and robust under geometric transformations, variation in the time duration of an event, and low resolution of images.

A recently proposed detection method, anti-faces [14], is shown to be very effective in the case of rich image collections which contain images under different geometric transformations such as scale, rotations, and projective distortions. Here we extend the method to image sequences. Consequently, the event detection method inherits the speed and robustness to geometrical distortions, which are the strongest features of anti-faces. We show that the temporal domain can be incorporated into anti-faces scheme as a third dimension. This implies that the extended method will be also robust to scale changes in time (different time durations). Specifically, change of speed (duration) in a time sequence can be seen as an affine transformation in three dimensions: $f(\mathbf{x},\mathbf{y},\mathbf{t}) \rightarrow f(\mathbf{x},\mathbf{y},\mathbf{at})$. Since the anti-face method works well for two-dimensional (2-D) affine transformation, it is reasonable to assume that its extension performs well for three-dimensional (3-D) affine transformations.

Manuscript received February 7, 2002; revised July 15, 2003. This work was supported by the Israeli Ministry of Scienceunder Grant 1229. This paper was recommended by Associate Editor Q. Tian.

The authors are with the Department of Computer Science, University of Haifa, Haifa 31905, Israel (e-mail: rita@research.nj.nec.com; dkeren@cs.haifa.ac.il).

Digital Object Identifier 10.1109/TCSVT.2004.825530

The algorithm proceeds by stacking the individual frames into spatio-temporal templates, and the detection is performed on these templates. The detection is done by applying detectors which are designed to yield small results on the sought event and large results on a "random" event. The detectors have the added bonus that they act in an independent manner, so that their false alarms are uncorrelated; consequently, the percentage of false alarms exponentially decreases in the number of detectors. This leads to a very fast detection algorithm, requiring only a small number of convolutions between the detectors and video sequence (viewed as vectors).

The proposed algorithm is able to discriminate the desired event from arbitrary "natural" sequences, and the "nonevents" are not restricted to a small predetermined training set of negative examples. This greatly simplifies the computation of the detectors—no database of negative examples is required—and also makes the detection more general.

We present three examples of applications that demonstrate a wide applicability of the proposed method. First, we show a synthetic example of rotating objects which demonstrates that, given a proper training, our method is robust to rotation, scale, and speed of the event. By proper training we mean that the training set must homogeneously sample the sought class of events. For example, if we want to detect variations in scale, samples from the desired range of scales should be present in the training set. As a result, the size of the training set will grow with the complexity of the class. However, the efficiency of the detectors is hardly affected by the complexity of the class. This was shown empirically in [14] by comparing the performance of the anti-faces with the eigenface method [25] on image classes of increasing complexity (different geometrical distortions). Experience has shown that, while the dimension of the face space in the eigenface method increased rapidly with the class complexity, there was almost no change in number of anti-face detectors required for correct detection.

The next experiment shows an application for visual speech recognition, which demonstrates that, even using low-resolution images, the proposed method is capable of discriminating a given word from very similar words. The last example is a feature-based application where we recognize the motion of features in a video sequence.

A. Previous Work

It is commonly accepted to divide the area of event detection into two parts: human action recognition and general motion-based recognition. Most of the approaches for understanding human actions require the existence of features which can be extracted from each frame of the image sequence, and then action recognition is performed on those features. Some of these techniques construct a 3-D body model [24], [10], [13],

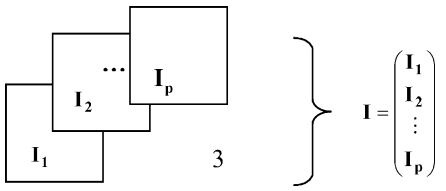


Fig. 1. Stacking frames. The sought event is represented by the sequence of video frames extending from I_1 to I_p.

[27], [29], and some compute image measurements and apply temporal models to interpret the results [16], [12], [21], [3], [5].

Other related work focuses on direct motion recognition [28], [19], [6], [1], [8]. One of the interesting recently proposed venues of research is the modeling of actions by basic flow fields, estimated by principal component analysis (PCA) from training sequences [1], [2]. The obvious difficulty with such an approach is that computing optical flow is nonrobust, which can affect recognition results. This is remedied by introducing robust estimation techniques.

The research proposed in [17] for lip reading is relevant to our approach, as it also uses frame stacks. The sequence of images of a spoken letter was taken as a 3-D template, where the third dimension is time. The authors extended the eigenface technique [25] to detect sequences of spoken letters.

The lip reading task was also studied in [4], [9], [11], [15], and [18]. Most of the techniques extract some features of the mouth area from each frame and then perform recognition by matching.

Another attempt to unify the spatial and temporal domains is offered in [6]. A "motion-history image" (MHI), which represents the motion at the corresponding spatial location in an image sequence, is built. This image captures only motion-related information. As mentioned by the authors, the weakness of such a technique is that in some cases it cannot discriminate between different motion directions, for instance, arm-waving in opposite directions. Another drawback is that the approach will fail in the case of motion with self-occlusion. In later work [7], the recognition framework was modified by computing local motion fields from the original MHI using a gradient-based motion pyramid and then characterizing an action by a polar histogram of motion orientations.

II. Event Detection in Video: The Gray-Level Domain

A naive approach for event detection is to perform object detection in each frame and then classify the object's motion. Due to low resolution of the video, illumination variability, and self-occlusion, this trivial solution may be limited. In addition, in event detection we are interested more in information existing "between the frames" than in the individual frames. Hence we use an entire sequence corresponding to the sought event as a 3-D template (Fig. 1) where two dimensions are spatial and the third dimension is temporal. For technical simplicity hereafter, we shall view such a template as a vector of size $N = \text{frame size} \times \text{sequence length}$.

A. Motivation

The recently proposed anti-face method [14] builds on the observation that natural images are usually smooth, in a sense that most of their energy is concentrated in the low frequencies. The same principle applies to video sequences: when viewed as a vector, the frame stack will usually be smooth. This follows from the fact that the change in natural video sequences is gradual; therefore, the function describing the variation in the temporal domain is smooth, as are the individual frames. As was formally proved in [14], **the absolute value of the inner product of two smooth vectors is, on the average, large**. We use this result to discriminate between a given event and "random" events. We build a set of filters (that we shall call detectors) that are smooth, thus their inner product with a random natural sequence of images will be large on average. We also require from the detectors to yield small inner products (hence "anti-sequences") with the events from the training set, in opposite to the large response that they yield for a "random" event. This provides a good discrimination between the positive class and any other event—not only negative examples.

More formally, if $\mathbf{d}$ is a candidate for a detector to the class T (collection of events that should be detected), suppose that not only is $|\langle \mathbf{d}, \mathbf{t} \rangle|$ small for $\mathbf{t} \in T$, but also that $\mathbf{d}$ is smooth. Then, if $\mathbf{y} \notin T$, there is a high probability that $|\langle \mathbf{d}, \mathbf{y} \rangle|$ will be large; this allows us to reject $\mathbf{y}$, that is, determine that it is not a member of T. Thus, a candidate event $\mathbf{y}$ will be rejected if, for some detector $\mathbf{d}$, $|\langle \mathbf{d}, \mathbf{y} \rangle|$ is larger than some $\mathbf{d}$-specific threshold.

B. Computing the Detectors

To find the first anti-sequence detector, $\mathbf{d}_1$, the following optimization problem should be solved (here we assume that T is a training set):

1) $\mathbf{d}_1$ is of unit norm.
2) $|\langle \mathbf{d}_1, \mathbf{t} \rangle|$ should be small for every image stack $\mathbf{t}$ in T. Note that every input is also normalized for the condition to make sense.
3) $\mathbf{d}_1$ should be as smooth as possible under the first and second constraints.

As was shown in [14], a commonly used image smoothness measure $\iint (\mathbf{I}_x^2 + \mathbf{I}_y^2) dxdy$ transformed to a frequency domain becomes a diagonal operator as follows:

$$S(\mathbf{I}) = \sum_{(k,l) \neq (0,0)}^{n} \left(k^2 + l^2\right) I^2(k,l) \tag{1}$$

where $I(k,l)$ are the discrete cosine transform (DCT) coefficients of an $n \times n$ image $\mathbf{I}$. In [14], (1) was used as a smoothness constraint for the detectors. We extend this constraint to image sequences by adding the third dimension associated with time; consequently, the smoothness constraint for the event detector $\mathbf{d}$ becomes

$$S(\mathrm{d}) = \sum_{(k,l,j) \neq (0,0,0)}^{n} \left(k^2 + l^2 + \left(\frac{j}{\alpha}\right)^2\right) \mathrm{d}^2(k,l,j) \tag{2}$$

where index j is associated with the temporal domain, $d(k,l,j)$ is the 3-D DCT transform of $\mathbf{d}$, and α is a scale factor adjusting

spatial and temporal "speeds." It was chosen so that the average of the absolute values of derivative in time is equal to the average of the absolute values of derivative in the spatial domain.

Since vectors are normalized to unit length, it is obvious that, by minimizing $S(\mathbf{d})$, we force the dominant values of the DCT transform $\{d(k,l,j)\}$ to concentrate in the small values of k,l,j— n other words, we make the detector $\mathbf{d}$ smooth.

The solution for the optimization problem defined above proceeds as follows. First, choose an appropriate value for $\max_{t\in T}|\langle \mathbf{d}_1, \mathbf{t}\rangle|$; experience has taught us that it doesn't matter much which value is used, as long as it is substantially smaller than the absolute value of the inner product of two random sequences. Next, minimize

$$\max_{t\in T}|\langle \mathbf{d}_1, \mathbf{t}\rangle| + \lambda S(\mathbf{d}_1) \tag{3}$$

and, using a binary search on λ, set it so that $\max_{t\in T}|\langle \mathbf{d}_1, \mathbf{t}\rangle| = M$.

We have used the Nelder–Mead method [20] for the optimization. The optimization is performed in the 3D DCT domain, and the inverse 3-D DCT of the optimum is the desired detector (note that the detection itself is carried out directly on the grey levels; the DCT domain is used only in the offline computation of the detectors).

There is a computational problem in the training stage, since stacking video frames results in very high dimensional vectors. Stacking one second of video with 25 fps and frame resolution of 100×60 produces a vector of dimension 150 000. One of the solutions is to compute only low frequencies of the detectors and pad the rest with zeros. However, once the detectors are recovered, their application is very fast.

C. Independent Detectors

Usually, a single detector is not sufficient to detect a given event with no false alarms; hence we apply several detectors which act *independently*, as was first proposed in [14]. Note that anti-face detectors are indeed independent under the assumption (made in [14]) that natural images behave according to Boltzman distribution which assigns higher probabilities to smoother images as follows:

$$\Pr(\mathbf{I}) \propto \exp(-S(\mathbf{I})). \tag{4}$$

From (1) and (4), it follows that the DCT coefficients of $\mathbf{I}$ are independent random Gaussian variables. Thus, if $\mathbf{f}_1$ and $\mathbf{f}_2$ are detectors, then $\mathbf{I_1} \rightarrow \langle \mathbf{I}, \mathbf{f_1}\rangle$ and $\mathbf{I_2} \rightarrow \langle \mathbf{I}, \mathbf{f_2}\rangle$ are also Gaussian random variables; this implies that if they are uncorrelated

$$\int_{R^{n\times n-1}} \langle \mathbf{I}, \mathbf{f}_1\rangle \langle \mathbf{I}, \mathbf{f}_2\rangle \Pr(\mathbf{I})\, d\mathbf{I} = \mathbf{0} \tag{5}$$

then they are also independent. As was shown in [14], (5) results in the following condition:

$$\sum_{(k,l)\neq(0,0)} \frac{f_1(k,l)\, f_2(k,l)}{k^2+l^2} = 0 \tag{6}$$

where f_1 and f_2 are the DCT coefficients of image detectors $\mathbf{f}_1$ and $\mathbf{f}_2$.

Since the smoothness assumption holds also for natural sequences,[1] we can extend the independence condition in (6) to video by adding temporal domain as a third dimension as follows:

$$\sum_{(k,l,j)\neq(0,0,0)} \frac{d_1(k,l,j)\, d_2(k,l,j)}{k^2+l^2+\left(\frac{j}{\alpha}\right)^2} = 0 \tag{7}$$

where d_1 and d_2 are the 3-D DCT coefficients of $\mathbf{d}_1$ and $\mathbf{d}_2$.

After $\mathbf{d}_1$ is found, it is straightforward to recover $\mathbf{d}_2$; the only difference is the additional condition in (7), and it is easy to incorporate this condition into the optimization scheme. The other detectors are found in a similar manner.

D. Detection Process

The detection process is very simple: an image sequence is classified as a given event, if and only if (iff) the absolute value of its inner product with each detector is smaller than some (detector specific) threshold. Only sequences which passed the threshold test imposed by the first detector are examined by the second detector, etc. Typically, the threshold was chosen as twice the maximum over the absolute values of the inner products of the given detector with the members of a training set for T. This factor of two allows detection not only of the members of the training set, but also sequences which are close to them.

The resulting detection algorithm is very fast; typically, $(1+\delta)N$ operations are required to classify a sequence of N pixels (when viewed as a vector), where $\delta < 0.5$.

III. Feature-Based Event Detection

The idea of frame stacking can also be applied to detect actions characterized by the movement of features (here, we used it to recognize symbols outlined by a laser pointer, and the feature was the pointers' image in any given frame). Each feature moving in a video sequence produces a curve $(x(t), y(t), t)$ in the spatio-temporal domain, which we shall call an "activity curve." The activity curve contains more than the geometric structure of the curve—it is also characterized by the speed and direction in which a point moves on the curve. Extracting the spatial positions of a feature in each frame and combining them to a single vector allows to apply the anti-sequence method to detection.

First, the sequence of triplets $(x(t), y(t), t)$ has to be converted to functions of one variable (t). The simplest method is to define the detection of an event as the detection of both $x(t)$ and $y(t)$. However, this simple approach is susceptible to symmetries in the spatio-temporal domain. For example, let us look at the case of a circle drawn counterclockwise; then, $x(t) = \cos(t)$, $y(t) = \sin(t)$. In the case of clockwise rotation, $x(t) = \cos(t)$, $y(t) = -\sin(t)$. Since the classification is based on the absolute values of inner products between the detectors and the templates, it will not be able to discriminate between a counterclockwise and a clockwise drawn circle. To remedy this problem, we modulate $x(t)$ and $y(t)$ by t. For example, we can

[1]By "natural sequences" we mean video sequences which, on the average, vary smoothly in space and time; this covers the very large majority of sequences which are of practical importance.

Fig. 2. COIL subset used in rotation sequence test.

define the corresponding curves as $x(t)+t$ and $y(t)+t$ (this is done, of course, both in the training and detection stages).

IV. Experimental Results

The following applications are studied: object rotation, visual speech recognition, and recognition of sketches outlined by a laser pointer. In all tests, the detection was very robust; on the average, there was a difference by a factor of more than ten in the detectors' response on the positive versus negative examples.

A. COIL Rotation Sequences

Two sets of experiments are presented. For the first test, we took 20 objects (Fig. 2) from the well-known COIL database[2] and simulated three types of video sequences: clockwise and counterclockwise rotation, and static. Then, for each object we built anti-sequences that discriminate clockwise rotation from other activities of the same object or other objects. The COIL database captures the objects in 5° rotation intervals. We created a training set from sequences of length five, capturing clockwise rotation with a 1–° phase between the sequences. For example, the first sequence consists of the respective object in 0°, 5°, 10°, 15°, and 20° angles; the second extends from 10° to 30°, and so on. In total, the training set included 35 sequences for each object. The test sequences for clockwise rotation were also created with a 10° phase between the sequences, but they started with 5°, then 15°, and so on. The counterclockwise rotation and static sequences of the same object were created with 5° phase. Hence, the experiment included 289 sequences, all of them disjoint from the training set. Ten anti-sequences were sufficient to discriminate the clockwise rotation of each object from the counterclockwise rotation of the same object and from any activity of the other items of the COIL database, with no misclassifications. The method produced 1% of false positives in static sequences. The reason for this is the short duration of the rotation sequences, since some of the objects hardly change during some of the sequences.

[2]Available. [Online]. http://www.cs.columbia.edu/CAVE/research/softlib/coil-100.html

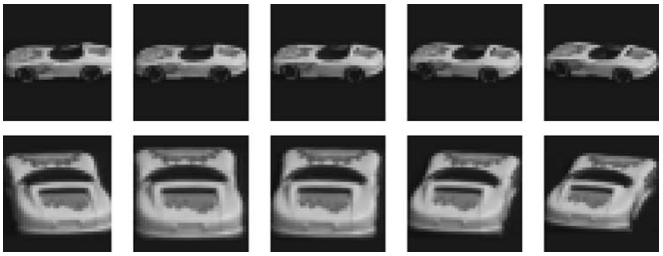


Fig. 3. Scale difference between different sequences.

Fig. 4. Anti-faces failed to discriminate between these similar objects, however, anti-sequences were able to discriminate between them.

The proposed method was applied to detect rotations in a wide range of velocities, extending from 5° to 20° between frames, with 4.8% of false alarms. This result demonstrates that the anti-sequence method is robust to variations in the event's duration. We should also mention that some of the tested objects change their scale a lot between the different sequences (see Fig. 3). However, the proposed method was able to handle both rotations and scales.

The next experiment was performed in order to compare the anti-sequence approach (viewed as an object detector) with anti-faces [14] applied to individual frames. The anti-face method required three to six detectors to discriminate an object from dissimilar items. To distinguish between the similar objects it usually required ten detectors, but it failed to discriminate between the objects in Fig. 4.

This experiment demonstrates that anti-sequences work well not only as event detectors, but they can also enhance *object detection*, which proves that the frame-stacking approach is more robust than what can be achieved by analyzing the individual frames. That is, applying anti-sequences as object detectors yields better results than applying object detection in all the individual frames.

In the second set of experiments, we addressed a more general problem: locate all instances of a particular object (a cup), performing clockwise rotation in the video sequence including the same cup performing rotations in both directions, a static cup, and several similar static and rotating objects. In this experiment, the search was performed in both the spatial and temporal dimensions. Fig. 5 shows fragments of the test sequence with the detection results marked by a white square around the detected image region. Six anti-sequences for the cup (Fig. 6) were sufficient to correctly detect its clockwise rotation.

B. Visual Speech Recognition

Next, we tested anti-sequences in recognition of spoken words. We captured 23 sequences of the word "psychology," uttered by a single speaker. Ten of them were used as a training set to generate the anti-sequences. The 13 remaining

Fig. 5. Fragment from the test sequence. Detection results for the five-frame sequences of clockwise cup rotation. The white squares mark the beginning of the detected sequences.

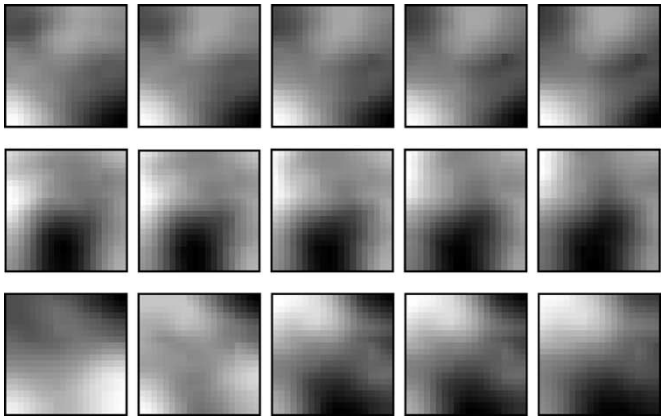


Fig. 6. First three cup anti-sequences. Note the smoothness in the spatial and temporal domains.

sequences and 20 other words were used in the recognition test. The acquired sequences had slight variations in global head movements and changes in the duration of the articulation. To reduce these undesirable variations, spatial alignment of the mouth position (see Section IV-B1) and time warping (see Section IV-B2) of the sequences were performed in the preprocessing step. Note that we used time warping because the training set contained only ten sequences, which are not enough to capture variability in articulation. However, we expect that a larger training set will allow for the removal of the time warping step from the algorithm. One of the "psychology" sequences was chosen as a reference (Fig. 7), and all of the test sequences were aligned and warped against it. The resulting sequences contained 26 images downsized and cropped to 24×16 pixels and centered around the lips. Since the mouth is symmetric in the x direction, we used only half of the images.

1) Spatial Alignment: Individual frames were aligned against the first frame of the reference sequence by extracting distinct features on the face and warping the frames by a corresponding rigid transformation. The features were the tip of the nose and two areas in the forehead; these were chosen as they do not change much while the person is talking.

2) Temporal Warping: Temporal warping was performed in the same way as described in [17]. The algorithm is based on the dynamic programming algorithm of Sakoe and Chiba [22].

Let W be the reference sequence with size N, and let A be an input sequence with size M that should be warped to size

Fig. 7. Reference sequence for the word "psychology."

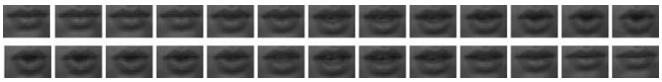


Fig. 8. The word "crocodile" warped to the length of the reference sequence.

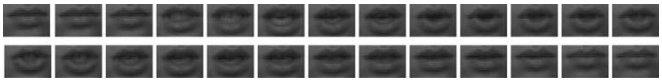


Fig. 9. The word "psychological" warped to the length of the reference sequence.

N. The warping algorithm uses the *DP-equation* in symmetric form with a slope constraint of 1 as follows:

$$g(i,j) = \min \begin{bmatrix} g(i-1,j-2) + 2d(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-2,j-1) + 2d(i-1,j) + d(i,j) \end{bmatrix}$$

where $d(i,j) = \|W_i - A_j\|$ is the distance from the ith element of the sequence W to the jth element of the sequence A. The initial conditions are

$$\begin{aligned} g(1,1) &= 2d(1,1) \\ d(i,j) &= \infty,\ g(i,j) = \infty, \qquad \text{for } i,j \le 0. \end{aligned}$$

The minimal argument chosen for the calculation of g at the point (i,j) defines the path from the previous point to the current one, thus creating a path from (1,1) to (M,N). Each point on the path indicates which frames from the input sequence match to frames in the reference sequence. In the case of two frames from the input sequence matching to one frame in the reference sequence, they are averaged to create a single frame. If one frame from the input sequence matches two frames from the reference sequence, it is duplicated. At the end of this process, the input sequences are warped to the size of the reference sequence.

3) Results: The experiment's goal was to recognize the word "psychology" in a test set that contained 13 instances of "psychology" which did not appear in the training set and 20 other words. The words tested for recognition were: "crocodile," "dinosaur," "encyclopedia," "transform," "integrable," "associative." "homomorphism," "leadership," "differential," "deodorant," "commutative," "anthropology," "trigonometry," "psychological," "anthology," "astrology," "cardiology," "dermatology," "genealogy," and "university." We have chosen the words such that some of them are totally different from "psychology" (like "crocodile" in Fig. 8), one word is very similar ("psychological" in Fig. 9), and the others have the same suffix ("ology") like the sought word (Fig. 10 shows the word "anthology").

Three anti-sequences (Fig. 11) sufficed to recognize all instances of "psychology" in the test set, with no misclassifications.

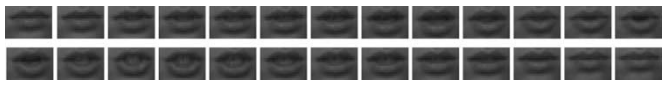


Fig. 10. The word "anthology" warped to the length of the reference sequence.

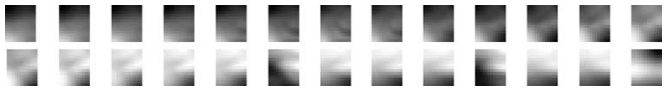


Fig. 11. First anti-sequence (of three) for the word "psychology." Since the mouth is symmetric, the anti-sequence frames are half the size of the images.

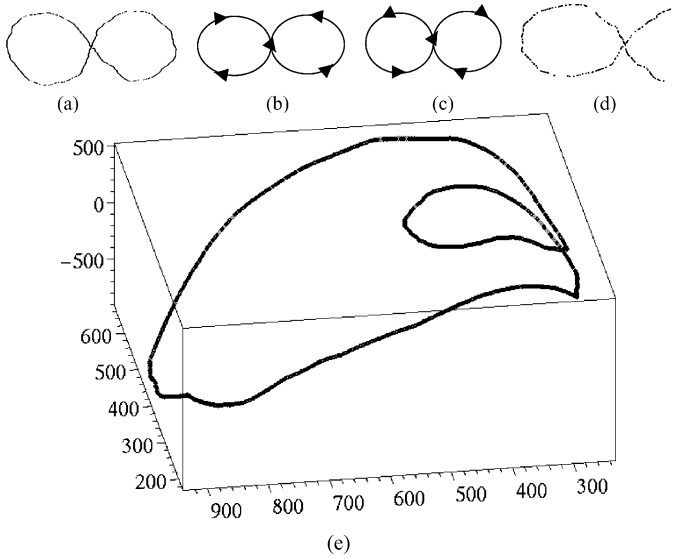


Fig. 12. (a) One of the "infinity" sequences used for training, (b) and (c) Schematic drawings of the infinity symbol traversed in two different directions. (d) The symbol α, one of the negative test examples. (e) A curve used for testing the detection scheme.

C. Symbol/Sketch Detection

This experiment concerned recognition of sketched symbols, and the results can also be applied to gesture detection. Various symbols were outlined with a laser pointer on a white background, and the process was captured on video. Then, the symbols were segmented from the background, and linear interpolation was applied in order to normalize them in space and time, which yielded a representation of each symbol as a vector with 200 points. Anti-sequence detectors were then constructed and applied to the one-dimensional vectors. The sought symbol was the infinity sign drawn at a certain order, shown in Fig. 12(b). The test set contained this symbol as well as the infinity sign drawn in a different direction, and other symbols (Fig. 12): α, β, γ, circle, square, and the digits 6, 8, 9. Two detectors sufficed to correctly detect the sought sketch with no false alarms, in all 50 tests performed. As shown in Fig. 11(a), detection was robust under both local and global deformations in the curve, which result from it being outlined by hand. Such deformations pose considerable difficulty for recognition methods which use differential invariants, as they are very susceptible to local distortions.

D. Sketch Detection in Three Dimensions

So far, we have discussed detection in which the training set contained sequences which approximate the sought sequence.

Fig. 13. Two basis views for the curve in Fig. 12.

Fig. 14. Three of the curve views that were detected by the algorithm.

In some cases, this assumption does not hold. Consider, for example, the problem of recognizing a sketched symbol which is 3-D (i.e., unlike the infinity symbol in Section IV-C, the sketch is not confined to a plane). This problem may arise, for example, in gesture detection; a certain feature, such as the tip of a finger, may outline a nonplanar curve. Suppose also that we do not know the angle from which the sketch was photographed. In principle, this problem could be alleviated by preparing a very large training set, which includes a dense sampling of the viewing sphere. However, this is very time consuming. The method presented here allows to detect different views using only a very small number of samples, by using the fact that a small number of views (*basis views*) span the entire view space. If a detector has a small inner product with these basis views, it will also have a small inner product with sequences that are spanned by them. Hence these sequences will also be detected.

The basic result we rely on appears in [26], where it is proved that for a transparent object in three dimensions, the x and y coordinates of all views can be expressed as a linear combination of two basis views. If a detector is suitable for these views, it will also detect other views (unless they are taken at angles in which the projection of the object is highly singular). In order to apply the method, corresponding points have to be chosen between the candidate view and the basis views. This was achieved by normalizing the curves to the same length, and using 200 evenly distributed points in both curves as the pairs of corresponding points. Since Euclidean length is not preserved between two projections of the same curve, we have used a measure of length which is invariant to affine transformations [23]. This was sufficient for a rather wide range of viewing angles.

We have tested these assumptions on the 3-D curve which is depicted in Fig. 12. The curve was sketched with a laser pointer on a curved piece of cardboard. The process was filmed by a video camera, and the curve's points were extracted from the individual frames (note that, as before, the curve has an order and rate of traversing associated with it). The two basis views are depicted in Fig. 13. The process was repeated with the camera positioned at other angles, and some of the resulting sketches which were detected are depicted in Fig. 14. In Fig. 15, some curves from the negative test set are shown.

Three detectors were required to successfully detect all the views of the curve in Fig. 12 that were tested, against ten other curves (negative examples), three of which are depicted in Fig. 15. To demonstrate the nature of the detection process, we have included in Fig. 16 the results of applying the first

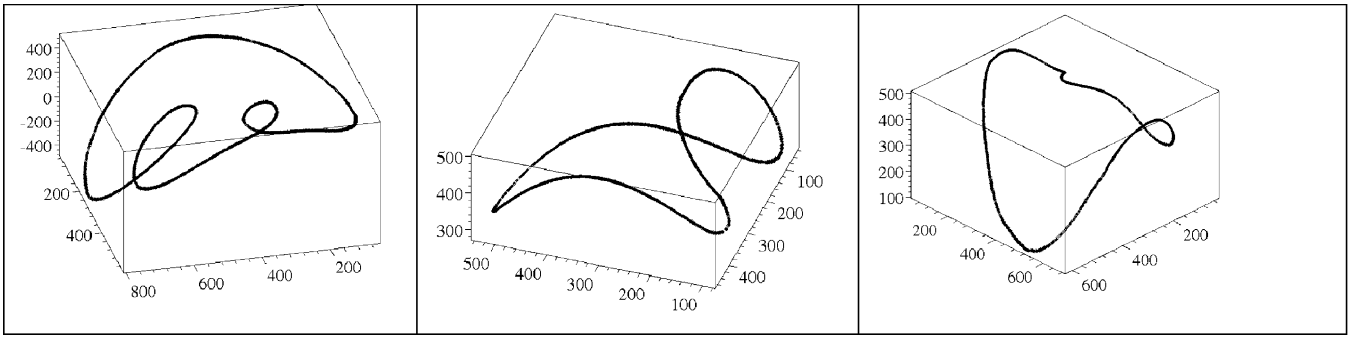
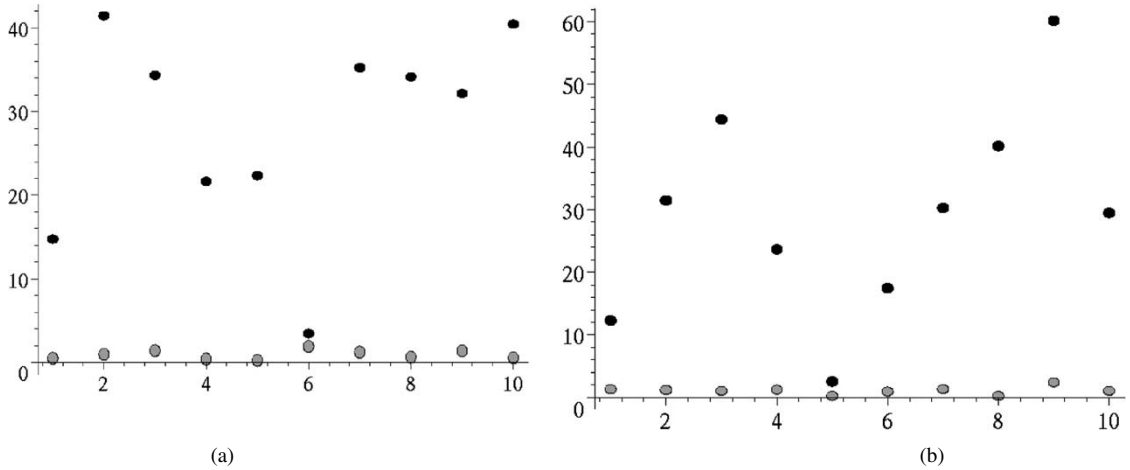


Fig. 15. Three curves from the negative test set.



(a) (b)

Fig. 16. Results of applying the (a) first and (b) second detectors to ten views of the sought activity curve (results depicted in gray) and the negative example in Fig. 15, center (results depicted in black). The hrizontal axis stands for view number, and the vertical axis for detector's output, scaled by 1000. While each detector admits one false alarm, the intersection of the sets of curves admitted by each of them successfully separates all of the views of the sought curve from all of the negative example's views.

two detectors to ten views of the sought curve, compared with the results for ten views of one of the negative examples (center curve in Fig. 15). Note that there is one false alarm for each detector (views 6, left, and 5, right), for which the corresponding detector yielded a small result, but no negative example passed the combined test of the two detectors. The threshold (scaled to the proportions in Fig. 16) was 8.0.

V. Conclusion

The "anti-face" method was extended to the time domain and used to detect events in video sequences. The algorithm was tested on sequences of rotating objects, and it was demonstrated that detection is more successful than separate detection in each frame and that it can detect rotations over a wide range of velocities. Another example was the detection of spoken words in a video sequence; the algorithm performed well, although the resolution was very low. The set of negative examples was not restricted, and contained words similar to the sought word. The algorithm was also applied to detect "activity curves", which correspond to sketched symbols in two and three dimensions. Using two "basis views," it was possible to successfully detect sketches in views that substantially differ from the training set views.

Future research will concentrate on expanding the method to detect "generic" activity (e.g., walking people).

References

[1] M. J. Black, D. J. Fleet, and Y. Yacoob, "Robustly estimating changes in image appearance," *Comput. Vis. Image Understanding*, vol. 78, pp. 8–31, 2000.

[2] M. J. Black, Y. Yacoob, A. D. Jepson, and D. J. Fleet, "Learning parameterized models of image motion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Puerto Rico, 1997, pp. 561–567.

[3] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Puerto Rico, 1997, pp. 568–574.

[4] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *Proc. IEEE Inter. Conf. Acoustics, Speech, and Signal Processing*, vol. II, Adelaide, Australia, 1994, pp. 669–672.

[5] S. Carlsson, "Recognizing walking people," in *Proc. Eur. Conf. Computer Vision*, 2000, pp. 472–486.

[6] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 257–267, Mar. 2001.

[7] J. W. Davis, "Hierarchical motion history images for recognition of human motion," in *Proc. IEEE Workshop on Detection and Recognition of Events in Video*, July 2001, pp. 39–46.

[8] I. A. Essa and A. P. Pentland, "Facial expression recognition using a dynamic model and motion energy," in *Proc. Int. Conf. Computer Vision*, Boston, MA, 1995, pp. 360–367.

[9] K. E. Finn and A. A. Montgomery, "Automatic optically-based recognition of speech," *Pattern Recognit. Lett.*, vol. 8, pp. 159–164, 1988.

[10] D. M. Gavrila and L. S. Davis, "3-D model-based tracking of humans in action: A multi-view approach," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1996, pp. 73–80.

[11] A. G. Goldschen, "Continuous automatic speech recognition by lipreading," Ph.D. dissertation, George Washington Univ.,, Washington, DC, 1993.

[12] D. Hogg, "Model-based vision: A program to see a walking person," *Image Vis. Comput.*, vol. 1, no. 1, pp. 5–20, 1983.

[13] I. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1996, pp. 81–87.

[14] D. Keren, M. Osadchy, and C. Gotsman, "Anti-faces: A novel, fast method for image detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 747–761, July 2001.

[15] M. Kirby, F. Weisser, and G. Dangelmayr, "A model problem in the representaion of digital image sequences," *Pattern Recognit.*, vol. 26, no. 1, pp. 63–73, 1993.

[16] M. E. Leventon and W. T. Freeman, *Bayesian Estimation of 3-D Human Motion from Image Sequence*: Mitsubishi Electric Research Lab, 1998, vol. TR-98–06.

[17] N. Li, S. Dettmer, and M. Shah, "Visually recognizing speech using eigensequences," in *Motion-Based Recognition*. Boston, MA: Kluwer, 1997, pp. 345–371.

[18] K. Mase and A. Pentland, *Lip Reading: Automatic Visual Recognition of Spoken Words*: M.I.T. Media Lab Vision Science, 1989, vol. TR 117.

[19] R. Polana and R. Nelson, "Low level recognition of human motion," in *Proc. IEEE Workshop Nonrigid and Articulated Motion*, Austin, TX, 1994, pp. 77–82.

[20] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes*. Cambridge, U.K.: Cambridge Univ. Press, 1986.

[21] J. M. Rehg and T. Kanade, "Model-based tracking of self-occluding articulated objects," in *Proc. Int. Conf. Computer Vision*, Boston, MA, 1995, pp. 612–617.

[22] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43–49, Feb. 1978.

[23] E. Salkowski, *Affine Differential Geometry*. Berlin, Germany: de Gruyter & Co., 1934.

[24] H. Sidenbladh, M. J. Black, and D. J. Fleet, "Stochastic tracking of 3D human figures using 2D image motion," in *Proc. Eur. Conf. Computer Vision*, 2000, pp. 702–718.

[25] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[26] S. Ullman and R. Basri, "Recognition by linear combinations of models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 992–1006, Oct. 1991.

[27] S. Watcher and H. H. Nagel, "Tracking persons in monocular image sequences," *Comput. Vis. Image Understanding*, vol. 74, no. 3, pp. 174–192, 1999.

[28] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," *Comput. Vis. Image Understanding*, vol. 73, no. 2, pp. 232–247, 1999.

[29] M. Yamamoto, A. Sato, S. Kawada, T. Kondo, and Y. Osaki, "Incremental tracking of human actions from multiple views," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998, pp. 2–7.