Anti-Sequences: Event Detection by Frame Stacking[1]

Margarita Osadchy Daniel Keren Yaniv Gal
Department of Computer Science
University of Haifa
Haifa 31905, Israel
e-mail: (gamer,dkeren)@cs.haifa.ac.il

Abstract

This paper presents a natural extension of the newly introduced "anti-face" method to event detection, both in the image and in the feature domains. In the case of the image domain (video sequences) we create spatio-temporal templates by stacking the video frames, 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.

The following applications of anti-sequences are presented: 1) Detection of an object under 3D rotations in a video sequence simulated from the COIL database, 2) Visual speech recognition of spoken words, and 3) Recognition of symbols sketched with a laser pointer.

The resulting detection algorithm is very fast, and is robust enough to work on small images. Also, it is capable of discriminating the desired event-template from arbitrary events, and not only events in a "negative training set".

1. Introduction

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 that there be individual features or properties that can be extracted from each frame of the image sequence, and then the recognition of action from the sequence of static features is performed. Some of these techniques build a 3D body model for action recognition [21],[9],[12],[23],[25], others compute image measurements and apply temporal models to interpret the results [15],[11],[19],[3],[5].

Other related work focuses on direct motion recognition. One of the interesting recently proposed directions is modeling of actions by basic flow fields, estimated by PCA from training sequences [1],[24],[18],[6], [1],[7]. The obvious difficulty in such an approach is that computing optical flow is highly non-robust; this can lead to poor recognition results.

In this paper, both image and feature domains are treated by creating spatio-temporal templates from the input sequence. We create these templates by stacking the video frames, and the detection is performed on the frame stack. The detection algorithm is a natural extension of the newly introduced *anti-face* method [13].

Relevant research is proposed in [16] for lip reading. The sequence of images of a spoken letter was taken as a 3D template where the third dimension is time. The authors extended the eigenfaces [22] technique to detect the sequences of spoken letters.

The lip reading task was also studied in [4],[8],[10][14], [17]. Most of the techniques extract the features of the mouth area from each frame and then perform the recognition by matching.

Another attempt to unify the spatial and temporal domains is offered in [6]. A "motion-history" image, which represents the motion at the corresponding spatial location in an image sequence, is created. This image captures from the entire sequence only the information related to the motion. As was mention by the authors the weakness of such a technique is that it cannot discriminate between the directions of motion, for instance arm-waving in opposite directions. Another drawback is that the approach will fail in the case of motion with self-occlusion.

The advantages of our approach relative to previous work are

1) The detection is very simple – convolution.
2) The detection is not restricted to a small predetermine set; for example, some previous papers concerning lip reading were restricted to collections such as some of the alphabet letters or the ten digits.

Experiments have produced encouraging results. The algorithm was able to discriminate the sought word (for example, "psychology") from similar words (for example, "psychological"), although the training set does not contain any negative examples.

[1] This work was supported by the Israeli Ministry of Science Grant no. 1229.

0-7695-1272-0/01 $10.00 © 2001 IEEE

1.1. Structure of the Paper

Section 2 starts with a short overview of anti-faces, and then extends it to event detection in both the image and feature domains by using frame stacks. Section 3 presents three different applications of anti-sequences: 1) detection of an object under 3D rotations in video sequence simulated from the COIL database, 2) lip reading, and 3) recognition of sketched symbols in the temporal domain.

2. Anti-Sequences

We present a fast algorithm for event detection in video sequences, which is a natural extension of anti-faces [13] to the temporal domain.

2.1. A Short Overview of Anti-Faces

Anti-faces [13] is a novel detection method, which works well in the case of a rich image collection – for instance, a frontal face under a large class of linear transformations, or 3D objects under different viewpoints. Call the collection of images, which should be detected, a *multi-template*. The detection problem is solved by sequentially applying very simple filters (or *detectors*), which act as inner products with a given image (viewed as vector) and satisfy the following conditions:

- The absolute values of their inner product with multi-template images are small.
- They are smooth, which results in the absolute values of their inner product with "random images" being large; this is the characteristic which enables the detectors to separate the multi-template from random images.
- They act in an independent manner, which implies that their false alarms are not correlated; hence, the false alarm rate decreases exponentially in the number of detectors.

The detection process is very simple: the image is classified as a member of the multi-template iff the absolute value of its inner product with each detector is smaller then some (detector specific) threshold. Only images which passed the threshold test imposed by the first detector are examined by the second detector, etc. 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. This allows detection not only of members of the training set, but also of images which are close to them.

This leads to a very fast detection algorithm. Typically, $(1+\delta)N$ operations are required to classify an N-pixel image, where $\delta < 0.5$.

To achieve invariance to the intensity of illumination, the images are normalized to zero mean and unit length.

2.2. Event Detection in the Image Domain

A naive approach to extend an object detection method to video sequences is to perform object detection in each frame and then classify the motion of the object. 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 take an entire sequence corresponding to an event as a template. The detection process will search for this template in the entire video sequence.

The basis for extending the anti-faces paradigm to event detection lies in the observation that, when viewed as a vector, the frame stack will usually be smooth. This follows from the fact that the change in natural video sequence is gradual, therefore the function describing the variation in the temporal domain is smooth.

We create the spatio-temporal templates by stacking frames (corresponding to the event) into one vector I [16]. Now, "anti-sequence detectors" are defined as vectors satisfying the conditions listed in section 2.1. The detectors are computed in the same way as was shown in [13] and the roughness measure is extended, with the third dimension being the temporal domain:

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

where $\tilde{I}(k,l,j)$ are the 3D DCT (Discrete Cosine Transform) coefficients of I and α is a scale factor adjusting spatial and temporal "speeds".

Usually a single detector is not sufficient to detect the event with no false alarms; hence we apply several detectors which act *independently*. This term is explained in [13]. The requirement that the detectors act independently implies the following condition:

$$\sum_{(k,l,j)\neq(0,0,0)}\frac{\tilde{D}_1(k,l,j)\tilde{D}_2(k,l,j)}{\left(k^2+l^2+\left(\frac{j}{\alpha}\right)^2\right)^{3/2}}=0$$

where $\tilde{D}_1$ and $\tilde{D}_2$ are the 3D DCT transforms of d_1 and d_2.

Once the detectors are found, the detection process proceeds in a similar manner to anti-faces.

The proposed algorithm for event detection is able to discriminate the desired event from arbitrary "natural" sequences, hence it is not restricted to a small predetermined (training) set.

In practice there is a computational problem in the preprocessing stage, since stacking video frames results in very high dimensional vectors. Stacking of one second of video with 25 fps and frame resolution of 100x60 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 extremely fast.

2.3. 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). Each feature moving in a video sequence produces a curve $(x(t), y(t), t)$ in the spatio-temporal domain. Extracting the spatial positions of a feature in each frame and combining them to a single vector allows to apply the anti-sequence method for 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 in 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 detector checks for the absolute values of inner products, it will not be able to discriminate between a counterclockwise and a clockwise drawn circles. To remedy this problem, we modulate $x(t)$ and $y(t)$ by t. For example, we define the corresponding curves as $x(t) + t$ and $y(t) + t$ (this is done, of course, both in the training and detection stages).

3. Experimental Results

In this section we demonstrate our approach on several different activities: object rotation, visual speech recognition, and detection of symbols outlined by laser pointer. In all tests the detection was very robust; there was a difference by a factor of more than ten in the detectors' response on the positive vs. negative examples.

3.1. COIL Rotation Sequences

In this section we describe two sets of experiments.
For the first test we took 20 objects (Figure 1) from the COIL database, and simulated three types of video sequences: clockwise and counterclockwise rotation, and static. Then, for each item of the COIL subset we built anti-sequences that discriminate an object rotating clockwise from other activities of the same object, or other objects. The COIL database contains the objects in 5-degree rotation steps. We created a training set from 5-frame sequences describing clockwise rotation of the object with a 10-degree phase between the sequences. In total, the training set included 35 sequences (for each tested object). The test sequences for the clockwise rotation of the same object were also created with 10-degree phase between the sequences, but they started from 5 degrees, then 15 and so on. The counterclockwise rotation and static sequences of the same object as well as for other COIL items were created with 5-degree phase, hence the experiment included 289 sequences, different 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 the sequences containing other items of the COIL database with no misclassifications. The method produced one percent of false positives in static sequences. The reason for this is the short duration of the rotation series, which in some angles are very similar to the static sequences.

The proposed method was applied to detect rotations in a wide range of velocities ranging from 5 to 20 degrees between frames, with a very low percentage of false alarms.

The next experiment was performed in order to compare the anti-sequence approach (viewed as an *object detector*) to anti-faces [13] applied to individual frames. The anti-face method required three to six detectors to discriminate an object from dissimilar items. To distinguish between similar objects it usually needed ten detectors, but in a few cases it failed to discriminate between the objects shown in Figure 2. This experiment showed that anti-sequences work well not only as event detector, they are also excellent as object detectors, which proves that the frame stacking approach is more robust than what can be achieved by analyzing the individual frames.

In the second set of experiments, we addressed a more general problem: locate all the 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 a cup was superimposed on a image consisting of 19 other objects. Figure 4 shows fragments of the test sequence with the detection results marked by a white square around the detected image region. Six anti-sequences of the cup (Figure 3) were sufficient to correctly detect its clockwise rotation.

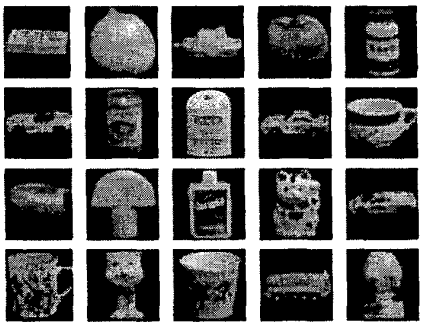

Figure 1: COIL subset for rotation sequence test.

Figure 2: The anti-face method failed to discriminate between these similar objects, however anti-sequences were able to discriminate between them.

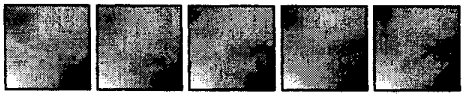

Figure 3: The first cup anti-sequence.

3.2. Visual Speech Recognition

In this section we have tested the 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 thirteen remaining sequences and other twenty words were used in the recognition test. The images were captured at a rate of 25 frames per second with 240x180 resolution. The acquired sequences had slight variations in global head movements and significant variations in the duration of the articulations. To reduce these undesirable variations, time warping of the sequences (see Sections 3.2.1) have been done in the preprocessing step. One of the "psychology" sequences was chosen as a reference (Figure 5) and all the sequences in this test were aligned and warped against it. The resulting sequences contained 26 images downsized and cropped to 24x16 pixels and centered around the lips. Since the mouth is symmetric in the x-direction, we used only half of the image, thus the image size was reduced to 12x16. Note that the time warping was performed to create a more uniform training set, since it has only 10 sequences and the variation in the articulation of the words was large. Increasing the size of the training set may eliminate the need in temporal warping of the sequences.

3.2.1. Temporal Warping

The temporal warping was performed in the same way as described in [16]. The algorithm is based on the Dynamic Programming Algorithm of Sakoe and Chiba [20].

The columns of each frame of a sequence are concatenated to form one vector, thus converting the image sequence to a sequence of vectors. Let W be the reference sequence with size N. Let A be an input sequence with a size M that should be warped to size N. The warping algorithm uses the *DP-equation* in symmetric form with the slope constraint of 1:

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

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

$$g(1,1) = 2d(1,1)$$

$$d(i,j) = \infty, \quad g(i,j) = \infty \quad \text{for } i,j \le 0$$

The minimal argument chosen for the calculation of g at the point (i,j) gives 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 case two frames from the input sequence match to one frame in the reference sequence, they are averaged to create a single frame. If one frame from the input sequence matches to two frames from the reference sequence, the frame is repeated. At the end of this process the input sequences are warped to the size of the reference sequence.

3.2.2. Results

The experiment's goal was to recognize the word "psychology" in a test set that contained thirteen instances of "psychology" which did not appear in training set, and twenty 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, university". We have chosen the words such that some of them are totally different from "psychology" (like "crocodile"), one word is very similar – "psychological" (Figure 6), and the others have the same suffix ("ology") like the sought word.

Three anti-sequences (Figure 7) were enough to recognize all instances of "psychology" in the test set with no misclassifications.

3.3. Symbol Detection

This experiment consisted of recognition of outlined symbols. The symbol to be detected was the infinity sign drawn in a certain order shown in Figure 8(b). The test set contained this symbol as well as the infinity sign drawn in a different order and other symbols (Figure 8): α, β, γ, circle, square, and the digits 6, 8 and 9. Two "anti-sequences" sequences sufficed to correctly detect the infinity sign drawn in a certain order, with no false alarms.

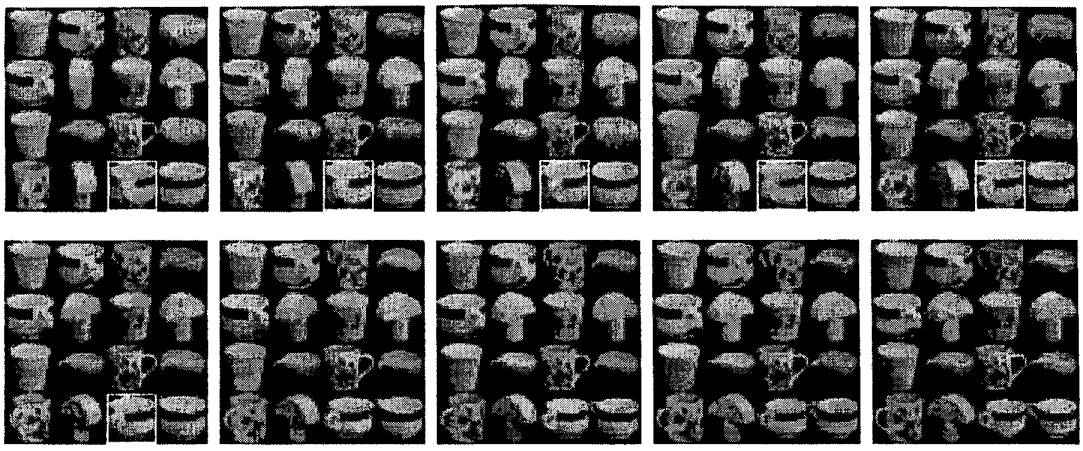

Figure 4: Fragment from the test sequence. Detection results of the 5-frame sequences of clockwise cup rotation. The white squares mark the beginning of the detected sequences.

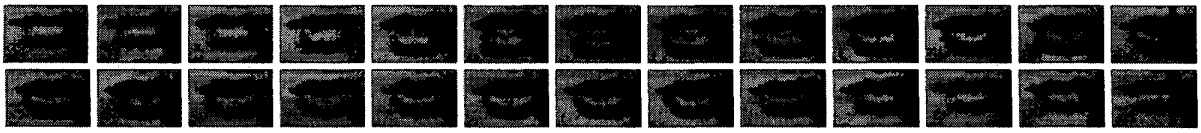

Figure 5: The reference sequence for the word "psychology".

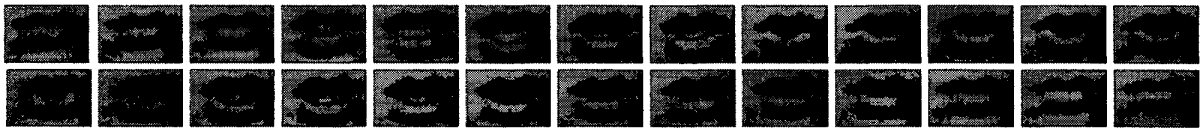

Figure 6: The word "psychological" warped to the length of the reference sequence.

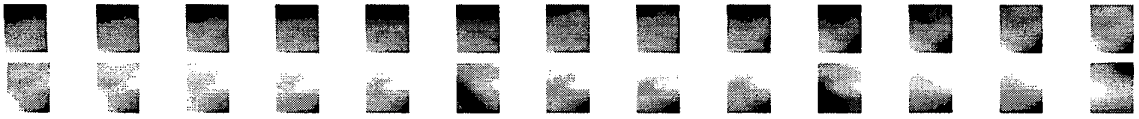

Figure 7: First anti-sequence (from three) for the word "psychology". Since a mouth is symmetrical, we built the anti-sequences only for a half of the mouth image.

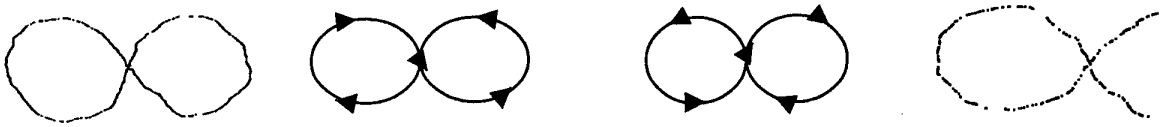

Figure 8: (a) one of the "infinity" sequences used for training; (b) and (c) schematic drawings of the infinity symbol traversed in two different directions; (d) the symbol α, one of the negative test examples.

References

[1] M.J. Black, D.J. Fleet, Y. Yacoob. "Robustly estimating changes in image appearance", *Computer Vision and 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," *IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[3] C. Bregler. "Learning and recognizing human dynamics in video sequences," *IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[4] C. Bregler and Y. Konig. "Eigenlips for robust speech recognition," *Proc. IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing*, 1994.

[5] S. Carlsson. "Recognizing walking people," *European Conf. on Computer Vision*, pp. 472-486, 2000.

[6] A.F. Bobick and J.W. Davis. "The recognition of human movement using temporal templates," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 23 (3), 2001.

[7] I.A. Essa and A. P. Pentland. "Facial expression recognition using a dynamic model and motion energy," *Proc. Int. Conf. on Comp Vision*, 1995.

[8] K.E. Finn and A.A. Montgomery. " Automatic optically-based recognition of speech," *Pattern Recognition Letters*, 8:159-164,1988.

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

[10] A.G. Goldschen. "*Continuous automatic speech recognition by lipreading*," PhD thesis, George Washington University, School of Engineering and Applied Science, 1993.

[11] D. Hogg. "Model-based vision: A program to see a walking person," *Image and Vision Computing*, 1(1):5-20, 1983.

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

[13] D. Keren, M. Osadchy, and C. Gotsman. "Anti-Faces: A novel, fast method for image detection," *IEEE Trans. on Patt. Anal. and Mach. Intelligence*.23(7),2001.

[14] M. Kirby, F. Weisser, and G. Dangelmayr. "A model problem in the representaion of ditital image sequences," *Pattern Recognition*, 26(1):63-73, 1993.

[15] M.E. Leventon and W.T. Freeman. "Bayesian estimation of 3D human motion from image sequence", TR-98-06, Mitsubishi Electric Research Lab, 1998.

[16] N. Li, S.Dettmer, and M. Shah. "Visually recognizing speech using eigensequences", *Motion-Based Recognition, Kluwer Academic Publishing*, pp. 345-371,1997.

[17] K. Mase and A. Pentland. "Lip reading: Automatic visual recognition of spoken words. TR 117, M.I.T. Media Lab Vision Science, 1989.

[18] R. Polana and R. Nelson. "Low level recognition of human motion," *IEEE Workshop on Nonrigid and Articulated Motion*, 1994.

[19] J.M Regh and T. Kanade. "Model-based tracking of self–occluding articulated objects," *Proc. Intern. Conf. on Computer Vision*, 1995.

[20] H. Sakoe and S.Chiba. "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustic, Speech, and Signal Processing*, ASSP-26(1):43-49, February 1978.

[21] H. Sidenbladh, M.J. Black, and D.J. Fleet. "Stochastic tracking of 3D human figures using 2D image motion," *European Conf. on Computer Vision*, pp. 702-718, 2000.

[22] M. Turk and A. Pentland. "Eigenfaces for recognition," *Journal of Cognitive Neuroscience* 3(1), 71-86,1991.

[23] S. Watcher and H.H. Nagel. "Tracking persons in monocular image sequences," *Comp. Vision and Image Understanding,"* vol. 74, no.3, pp. 174-192, 1999.

[24] Y. Yacoob and M.J. Black "Parameterized modeling and recognition of activities," Computer Vision and Image Understanding, vol 73, no. 2, pp 232-247, 1999.

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