CAPTURING THE HUMAN ACTION SEMANTICS USING A QUERY-BY-EXAMPLE

Anna Montesanto, Paola Baldassarri, A. F. Dragoni, G. Vallesi and P. Puliti
DEIT, Università Politecnica delle Marche, Italy

Keywords: Query by example, human action semantics, artificial intelligent systems, neural networks.

Abstract: The paper describes a method for extracting human action semantics in video's using queries-by-example. Here we consider the indexing and the matching problems of content-based human motion data retrieval. The query formulation is based on trajectories that may be easily built or extracted by following relevant points on a video, by a novice user too. The so realized trajectories contain high value of action semantics. The semantic schema is built by splitting a trajectory in time ordered sub-sequences that contain the features of extracted points. This kind of semantic representation allows reducing the search space dimensionality and, being human-oriented, allows a selective recognition of actions that are very similar among them. A neural network system analyzes the video semantic similarity, using a two-layer architecture of multilayer perceptrons, which is able to learn the semantic schema of the actions and to recognize them.

1 INTRODUCTION

In the literature several Content-Based Video Retrieval (CBVR) papers have been published. CVBR appears like a natural extension of Content-Based Image Retrieval (CBIR) and Content-Based Audio Retrieval systems (Marchand-Maillet, 2000). However there are a number of factors which should be dealt with when using video that are ignored when dealing with images. These factors are related to the temporal information available from a video document. The temporal information induces the concept of motion for the objects that are in the video. This information allows us to encode within the indexing of a video the behaviour of all objects throughout the document.

An important aspect to evaluate is the complexity of the query system because a video contains a greater amount of information than a textual description. In this study, we intend to supply a method of query where a user can specify the features of interest using a sketch pad or the whole video like example. A partial solution could be represented by a query-by-example: through the query a user can specify the characteristic of the video that he searches. There Behaviour understanding involves the analysis and recognition of motion patterns, and the production of high level description of actions and interactions (Hu et al., 2004). We can suppose that the understanding of the actions is related to the temporal classification of particular features and to the matching with an unknown sequence with previously learned sequences representatives of typical behaviours. Then a fundamental problem of behaviour understanding is to learn the reference behaviour sequences from training samples and to devise both training and matching problems. Some efforts have been made in this direction (Bobick & Davis, 2001; Venkatesh Babu & Ramakrishnan, 2004; Kang et al., 2004) and some methods of behaviour understanding are outlined in the following. Other authors (Bobick & Wilson, 1997) used Dynamic Time Warping (DTW) to match a test sequence to a deterministic sequence to recognize human gestures. Hidden Markow Models (HMM) consisted of training and classifications outperform DTW for undivided time series data and so are used to behaviour understanding (Brand & Kettnaker, 2000; Oliver et al., 2000). Time Delay Neural Networks (TDNN) has been used to hand gesture recognition and lip reading (Meier et al, 2000). The self-organising neural network can be applied for scenes where the type of object in motion is not known (Owens & Hunter, 2000; Sumpter & Bulpitt, 2000).

In this paper we focus the attention on the movement of a single object. In particular we intend to realize a system that is able to recognize a human movement. The process of human movement recognition is

Montesanto A., Baldassarri P., F. Dragoni A., Vallesi G. and Puliti P. (2008).
CAPTURING THE HUMAN ACTION SEMANTICS USING A QUERY-BY-EXAMPLE.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 356-363
DOI: 10.5220/0001932703560363

based on a schema in which the whole action is organized in a kind of macrostructure. A schema is a cognitive structure characterized by cases that are arranged in a hierarchical way, but they are also restructured on the base of the context (Rumelhart & Ortony, 1964). Moreover the schema also organizes the information that can be extracted in different levels. Our schema consisted of a temporal sequence subdivided in 10 steps. Three are the parameters characterizing the trajectory of the action: the (x,y) coordinates and the (v) instantaneous velocity. We suppose that these three spatio-temporal parameters constitute the minimal information representative of the action semantics. The choice of these variables reduces the dimensionality of the space of searching and then respects a user-oriented representation.
Each schema of movement used to recognize an action can represent a class of movement and different movements can belong to the same class.
In order to recognize a human movement a visual system has not to passively record local motion since they are ambiguous (kinematics indetermination), but the local signals have to be integrated in a coherent local motion. In a classic demonstration (Johansson, 1973) were created displays of human movement by attaching lights to the major joint of human models. The models were recorded so that only the point lights were visible to the observers. So the observers had vivid impressions of the human figures, although the images contained only few bright points. So, the movement is enough to extract meaningful information of 3D shapes.

In (Runeson, 1994) it is suggested that an accurate appraisal of kinematics aspects of the movement allows us understanding its dynamic aspects. It is important to estimate the active forces that cause the movement. A very important aspect is the co-variation of kinematics and geometry that is well exemplified from a lot of study about the hand gestures. They satisfy some principles of invariance that characterize and distinguish them from the movements of mechanical type (Viviani & Schneider, 1991).

So we can affirm that the semantics of the action corresponds to the human perception of the co-variation of kinematics and geometry. By departing from this point of view, we have built a system that concurs with the user to formulate its query using a representation that implicitly corresponds to the semantic-cognitive schema of the humans. Such modality of query formulation goes behind the problem of the query by example in the video, since it is not necessary a particular training to use this system. Moreover the same system can easily succeed to extract the semantic aspects of the action.

Our recognition system has also to form a base of knowledge for correctly classify, but above all it must be sufficiently resistant to the noise. This last could be the variation of the point of view from which the action has been recorded. A particularly suitable methodology uses the artificial neural nets which are trained through a learning process on known scenes.

2 SYSTEM OF QUERY-BY-EXAMPLE IN VIDEO

We realised 4 sub-systems, communicating between them through an exchange of information (memorisation and recovery) carried out using a database.

2.1 Learning Phase

The proposed system has to recognize and to codify the actions of the query. Therefore, in a preliminary phase the videos are inserted in a database, and so they represent the pattern of training. For that, the manager of the system (an expert user) determines the number n of videos for the training of the system for the recognition of the actions and the number τ of typologies of actions. From that it is created a table in which the list of the actions encoded through a binary code is represented. So, this table contains the list of actions used for the learning. After carried out the tracking of the point, its trajectory is inserted in the database pointing to the video from which come from. The video is identified by a series of features: the code of the typology of the action and the coordinates (x,y) of the points. At this point, it is possible to supply a feedback for the user designing the video trajectory superimposes to the first frame. The coordinates of the point are sent to the "analysis module" which it calculates the third feature characterizing the trajectory is the instantaneous velocity (v). All the information concerning the video (the binary code of the action, the x,y coordinates and the velocity v) are stored in a file.

The data of the file are the inputs of the "recognition module". In this module start the phase of training of the neural architecture. At the end of the learning the weights vectors of the neural networks architecture are trained in order to allow the system to recognize and to classify a particular

action during the testing or the generalization phase. These vectors are added to the characteristics of the trajectory of the action in the database.

2.2 Generalization Phase

After the training of the system we can use as query a video that the system has never seen before. Such video is recalled from the system of tracking. The novice user clicks on the frame evidencing the interest point of which made the tracking and the trajectory is inserted in the database with the pointer to the file of the video. Specific features characterize the trajectory of the action: the coordinates of the points belonging (but it has not the typology of action because the novice user has not to necessarily express it inside the query). Now is supplied a feedback for the user: it is draw the video trajectory superimposes to the first frame.

The coordinates of the points of the trajectory are sent to the analysis system which it calculates the instantaneous velocity v. The parameters are the inputs to the "recognition module" that classifies the trajectory indicating the class of belonging of the analysed action.

Moreover this information is sent to the system of matching that collects all the trajectories that have the same class vector of belonging to the database.

The results are shown to the user as the first frame with over drawn the trajectory of the action; such frame is clickable and let's starts the video.

3 THE TRACKING ALGORITHM

Starting from a video, this algorithm made the tracking of the points chosen by the user or of the point that it retains optimal to be tracked. Then it extract the coordinates x and y of such points.

The algorithm used to realize the tracking is based on the pyramidal Lucas-Kanade technique (Bouget, 1999) for its performances of robustness and accuracy. Concerning the accuracy we need small integration window for not smoothing the details of the image. While the robustness concerns the power of the algorithm related to the light variations, the image dimensions and the motion. So the robustness need large image.

In this method the pyramidal representation is used because it is simpler to follow wide movements of pixel, that they are wider than the window of integration at the main image level, while they are confined in the same window at the lower level. Resuming it is important a pyramidal representation of the image, in which in the high level we obtain a small image that, allows precise comparison among the pixels of two consecutive frames. While in the lower level we obtain a large image that allows maintaining the macroscopic differences of the movements.

For each analysis level the Lucas-Kanade algorithm halves the image resolution. The Lucas-Kanade algorithm considers two frames at a time and for each pixel of the first image it finds the corresponding in the successive image. So, considering different image resolution we intend to minimize the "Residual Error" function $\epsilon(\cdot)$, that derives from the summation of all the motion vectors related to the iteration level:

$$\epsilon(\mathbf{d}) = \epsilon(d_x, d_y) = \sum_{x=u_x-\omega_x}^{u_x+\omega_x} \sum_{y=u_y-\omega_y}^{u_y+\omega_y} (I(x,y) - J(x+d_x, y+d_y))^2 . \quad (1)$$

where I and J represent two frames, while ω_x e ω_y indicates the dimensions of the window of integration. These parameters have to be appropriately chosen in order to guarantee the performances of robustness and accuracy typical of the algorithm.

Having the (x,y) point in a particular frame, we can determine the corresponding point in the successive frame. The (x,y,v) tern is stored in a database (Yoon et al., 2001), where v is the instantaneous velocity. The matrix contains the spatio-temporal minimal information representing the semantic of the action. The semantic is defined by the way in which the motion in time develops.

In particular we can affirm that the dynamical constraints render perceptively unique the action, even if the geometry of its trajectory is very similar among different actions. The (x,y,v) tern allows us studying the motion considering the implicated forces, but only if we extend the information on the complete temporal sequence.

Such algorithm, moreover, has been optimised through a procedure that allows the user choosing the optimal points. They are those points that carry with greater probability to good results. The computation of the gradient of the brightness function has been implemented in such way that the user can choose between the best points that have a gradient value upper to a fixed threshold. To such purpose, during the recording of the movies the subject wore a bracelet with a white marker. So in correspondence of the wrist we have a strong gradient of brightness and therefore a greater precision in phase of tracking.

4 ALGORITHM OF ANALYSIS AND INDEXING

This algorithm read the trajectory coordinates and draws the shape of it. Also using those coordinates it computes all the necessary features. In particular it extracts the instantaneous velocity of every point of the trajectory. It uses the data store in the DB in order to process them or to visualise them. The first phase of the algorithm assigns to a vector the coordinates of the points individualised from tracking (abscissa and ordinate); while in the successive phase the trajectory is represented. It is applied, therefore the algorithm of chain-coding, that it chases the points finds realising one chain codified. In such phase we start from the first stored point, centred in a matrix 3x3, and we look for around to it, the successive point pertaining to the trajectory. Once characterised the successive point, moves the matrix and restart the search. In this way, starting from the first point, the chain is automatically generated and characterises the dynamic of the movie. Once stored the trajectory, is carried out the analysis and therefore the extraction of its characteristic parameters.

The interface visualise an image that represents the mosaic of the considered movie, used like background and visual reference. On such image the program draws the codified trajectory and on the same window the extracted parameters are described.

5 ARCHITECTURE OF PATTERN RECOGNITION SYSTEM

We realized a system consisted of two layers architecture. The first layer consists of 10 perceptrons, while the second layer consists of only one perceptron. The system receives in input the features extracted from the video sequences and gives in output which is the kind of action that a subject performed. The three actions (eating, drinking and smoking) are codified as subsequently explained.

We subdivided the trajectory of each action in 10 steps each consisted of 8 frames. This subdivision is based on the hypothesis that an action is recognized through an initially general scheme. The scheme is defined during the development of the action itself.

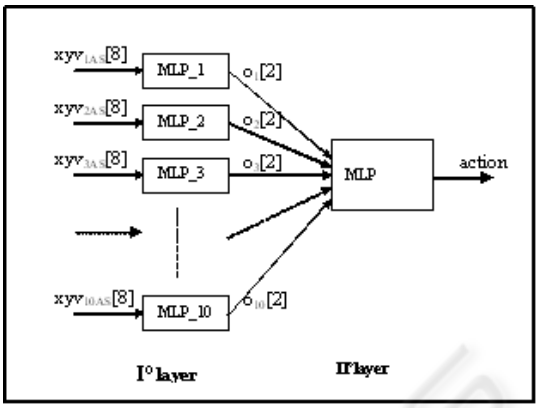

Figure 1: The architecture of the neural system.

The figure 1 shows architecture consisted of two neural networks levels. In the first level we have 10 Multilayer Perceptrons (MLP) (Rumelhart et al., 1986) each of them has 24 nodes in the input layer. Each node of the input layer receives the values xyv for all the subjects and for each action. The vector xyv_{iAS} identifies the elements of the trajectory, where the index i (that goes from 1 to 10) indicates the number of steps in which the trajectory is subdivided. The index A = E,D,S identifies the three action, and the index S (that goes from 1 to N) identifies a particular subject (N is the number of the subjects). So the dimension of the input for each network of the first layer is equals to $8\times3\times N$.

The hidden layer consists of 12 nodes, while the output layer consists of only 2 nodes. The vector $o_k[2]$ identifies the output of this last layer, where k goes from 1 to 10 (number of steps in which the sequence is divided). The vector o_k consists of two elements, these identify a specific action according to the following code: [0,1] for the eating action, [1,0] for the drinking action and [1,1] for the smoking action.

The second level consists of only one MLP which receives as input the outputs of the previous 10 neural networks. So in the input layer it has 20 nodes, in the hidden layer it has 5 nodes and finally in the output layer it has 2 nodes, which encode the kind of action.

6 TEST ON HUMAN ACTION RECOGNITION

We want to see how sensitive was the neural system to recognize sequences of action very similar one another, in particular from the geometrical point of view. The experiment is made so to study the

answers invariance changing the point of record of the camera. We can observe events that are equal as regard the trajectory and are different as regard the semantics: these can hide in practice their actual meaning in small differences.

6.1 Methodology

In order to try the effectiveness of the parameters in an unambiguous recognition of the action, the motion of the wrist of subjects while they are eating, drinking and smoking was recorded. The movies have been recorded always to the same distance, using a fixed experimental setting. This has made so that the trajectories were much similar between them. The recorded sample consists of 25 subjects.

We have chosen simple actions to carry out and to study: Eat, Drink and Smoke.

In each action, the movement is always of the arm, always to leave from the bottom to the up, until arriving and stopping themselves on the mouth, and then from here, to return behind until stopping itself, caught up the desk newly.

The technical apparatus consists of two analogical television cameras: one of these was in a frontal position, whilst the other was in a lateral position respect to the subject. There are also one desk, one chair, and finally cracker, water and cigarettes. Above all were fixed the relevant positions, so that the features unchanged. We arranged the desk and the position of the chair relatively to it.

Then it has been taken the centre of the desk and in connection to this, one camera is blocked to 5 meters of distance and 53° approximately of angle-shot as regards the lateral position, whilst the second is blocked to 5 meters and concerns the frontal position. It is moreover fixed on the table the start-position of the objects (cracker, water, and cigarettes), so the gesture had a precise fixed point of reference.

Finally, in order to obtain a metric calibration of the digital images, in order to have a correlation between the real distances and the resolution in pixel after the digitalisation, a graded bar has been placed (it was 30 centimetres long).

During the recording, each subject wore a bracelet with two white markers, obtaining a high gradient of brightness in the point of the body that mainly characterises the movement.

To extract the interesting data from the movies we digitised them (card of acquisition ATI All in Wonder 128), to 20 frame for second with a resolution of 320x240 pixels and with a colour depth of 24 bits. Then we transform them in grey-scale levels. We compressed the video sequences through the algorithm of Run Length Encoding: that replaces the sequences of identical pixel with the indication of the number of epochs in which the pixel is repeated, followed from the value of the pixel itself.

Each of the 50 movies (25 for each camera) has been decomposed in 3 shorter movies (correspondent to a particular action), every one of which has been then analysed singularly.

7 EXPERIMENTAL RESULTS

During the learning, in all the experiments in order to achieve comparable results we fixed the same number of iteration epochs and the same learning rate. In particular the number of iterations was fixed to 15000 epochs and the learning rate was $\eta = 0.1$.

Remembering that each of the three actions was subdivided in a sequence of ten steps, we determined the mean square error for each step after 15000 iterations, or rather in the end of the learning. The figure 2 shows the trend of the mean square error after 15000 iterations: the black line refers to the learning by the three actions of 25 subjects of lateral view, whilst the grey line refers to the learning by the three actions of 25 subjects of frontal view.

Considering the training with the 25 subjects of the lateral view, the average error calculates on all the sequence is 0.04. If we divide the sequence in two phases (going and return): the average error of going is 0.006, whist the average error of the return is 0.08.

Moreover training the networks with the 25 subjects of the frontal view the average error is 0.03. Dividing the sequence in two phases the average error for the going is 0.01, whilst the error is 0.05 for the return phase. These values are understandable if we consider the “kind” of action. In order to have more possible similar actions, all the subjects started the movement from a fixed point, caught up the mouth and then returned in the fixed starting point. Initially the action is loaded of meaning both for the aim linked to the action and for the object grasped by the end. The heaviness of a water glass is certainly different from a cigarette. So, in the going the actions are quite different, instead in the return they show a conjunction of meaning: all the movement have only to reach a fixed point. The consequence is the increase of the average error, so the not ambiguous recognition of a particular action is closely related to the development of the sequence.

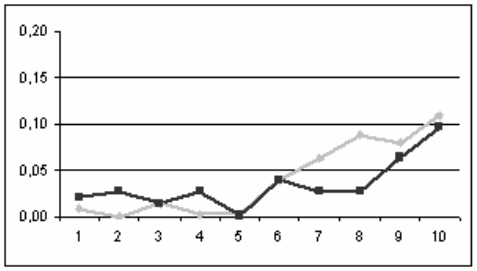

Figure 2: Trend of the mean square errors for each net of the first layer after 15000 iterations (black line = training with 25 subjects of lateral view; grey line = training with 25 subjects of frontal view).

7.1 Experiment 1: View Point Invariance

In the first experiment we want to demonstrate the invariance from the point of view. The aim is to verify that a particular action (that is eating, drinking and smoking) of a subject is recognizable apart from the point of view. In fact, in this example we considered both the frontal view and the lateral view of the same subject in the act of doing one of three actions previously described.

The system was trained using the lateral shots of 25 subjects during the development of the three actions. During the testing we used only the frontal shots of the same 25 subjects in the act of doing the same three actions. We have to remember that the two shots took place contemporaneously from two cameras one in a lateral positions and the other in a frontal position as regards the subjects.

The training using the parameters (x,y and v) extracted by the lateral view and the successive testing using the same parameters extracted by the frontal view allows evaluating the independence from the point of view.

This implies that the system is able to recognize an action done by a subject independently from the observation position.

Table 1: Confusion Matrix.

	EAT	DRINK	SMOKE	
EAT	19	0	6	25
DRINK	0	18	7	25
SMOKE	3	4	18	25

The table 1 shows the confusion matrix of the test results. The first rows and the first column indicate the three different actions done by the 25 subjects. The last column indicates the total number of the subjects that they do a specific action. The main diagonal shows the total number of actions correctly recognized. For example as regards the eating action (EAT), the system correctly recognizes 19 subjects out of 25 subjects during the execution of the eating action. As we can see in confusion matrix, both the eating action and the drinking action are mistaken with the smoking action. In other words, the system misses to recognize both some eating actions and some drinking actions identifying them with smoking. Finally, the last row of the table 1 shows that the smoking action was identified 3 times (12%) with the eating action and 4 times (16%) with the drinking one.

The rates of recognition are comparable between them. The rate of recognition is 76% for the eating action, 72% for the drinking action and finally 72% for the smoking action. For the first experiment we obtained an average rate of recognition equals to 73.3%. The experimental results show that the invariance from the point of view is rather well respected.

7.2 Experiment 2: Subject Invariance

The second experiment intends to demonstrate the performances of the system concerning the invariance from the subject. In other words, we want to demonstrate that the system is able to correctly recognize an action independently from the subject, who makes this. For this purpose, the system was trained with 15 subject of the frontal view in act of doing the three actions, while for the test we considered the others 10 subjects of the same frontal view. The correct recognition of a particular action implies that the development of an action is not determined by the particular subject involved, but from the characteristics implied in the action itself.

Table 2: Confusion Matrix.

	EAT	DRINK	SMOKE	
EAT	10	0	0	10
DRINK	0	9	1	10
SMOKE	1	0	9	10

As for the first experiment, we introduced the confusion matrix of the actions the system correctly recognizes (table 2). Each of the 10 subjects carries out the three actions (eating, drinking and smoking), in total 30 actions.

The results in the table 2 put in evidence that the system correctly recognizes the eating action carried out by the 10 subjects of the testing phase: so the eating action is recognized with a rate of 100% (corresponding to 10 out of 10). The drinking and

the smoking actions are also recognized with a rather high rate: the system correctly recognizes both the drinking action and the smoking 9 out of 10 times.

We obtained an average rate of recognition equals to 93.3%. The three actions are correctly recognized independently by the subject who makes the action.

7.3 Experiment 3: View Point and Subject Invariance

The last experiment intends to combine the two previous experiments. In fact, in this part we want to demonstrate that the system is able to recognize the three actions (eating, drinking and smoking) changing both the point of view and the subject who makes the action. For this purpose, the system is trained using the lateral view of the 15 subjects during their movement. While the testing procedure is executed using the frontal view of the other 10 subjects in act of doing the same actions. The training using the parameters (x,y, and v) extracted by the lateral view of 15 subjects and the successive test using the same parameters extracted by the frontal view of the other 10 subjects intend to evaluate the independence of the system both from the point of view and from the subject. This implies the ability of the system to recognize an action independently from the position of observation of the subject and independently from the subject himself.

Referring to the usual representation of the results, we analyze the confusion matrix in the table 3.

The confusion matrix put in evidence that the three actions are rather well recognized. Observing the main diagonal we can see that the eating action is recognized 8 out of 10 times, the drinking action 9 out of 10 times and the smoking action 7 out of 10 times.

Table 3: Confusion Matrix.

	EAT	DRINK	SMOKE	
EAT	8	0	2	10
DRINK	0	9	1	10
SMOKE	1	2	7	10

As for the first experiment both the eating action and the drinking action are confused with the smoking action. The smoking action was identified for a 10% with the eating action and for a 20% with the drinking.

We can determine an average rate of recognition equals to 80%. In the last experiment we also obtained rather encouraging results, having a high rate of recognition for all the three actions.

8 CONCLUSIONS

The implemented neural architecture learns to recognise the kinematics and the dynamic aspects of actions concerning the trajectories but also the velocity variations. The experimental results showed the real possibility of building a system that allows a semantic recognition of an action, taking advantage from the action recognition modalities of the human. The lightness of information joined to the generalisation abilities of the neural networks, make the system extremely effective. The typology of action is recognized also if we observe a subject from two different points of view (lateral and frontal view), and so the invariance of the system as regard the point of view is rather well respected. Moreover the results demonstrated that the system recognized a specific movement performed by a different subject from those one learning during the training and observed from the same point of view (lateral view). So we can conclude that our system catch also the subject invariance. Finally we test the system invariance changing both the point of view and the subject who makes the action, obtaining high value of recognition rate.

In all the experiments the recognition of each action has enough high values, but they could be better in the case in which every networks of the first layer will be trained with a number of iterations that fit with the minimum square error. In fact, in this work we stopped the simulation after 15000 steps to achieve comparable results evaluating the suitability of particular chosen knowledge representation. The experimental evidences show that the three spatio-temporal parameters (x, y, v) constitute the minimal information representative of the action semantics. The neural networks architecture allows us to have a smart classification of the part of sequence in which there is a great amount of different semantics.

One of the possible future developments of the work can be using the system as recogniser of the behaviour of a given subject that may be used in surveillance applications. The study realized in this work could be applied in numerous and different environments but with an only common feature: the necessity of understanding not “what there is in a scene?", but “what is happening?”. For instance we

could develop safety systems able to recognise anomalous and dangerous situations by automatic extraction of scene semantics. This represents an evident advantage of efficiency.

REFERENCES

Bobick A. and Davis J., 2001 "The Recognition of Human Movement Using Temporal Templates," IEEE Transactions on Pattern Recognition and Machine Intelligent, vol.23, pp. 257–267.

Bobick A.F. and Wilson A.D., 1997 "A State-Based Technique to the Representation and Recognition of Gesture", IEEE Transactions on Pattern Analysis and Machine Intelligent, Vol.19, pp.1325-1337.

Bouget J.Y., 1999 "Pyramidal Implementation of the Lucas Kanade Feature Tracker. Description of the Algorithm", Intel Corporation, internal report.

Brand M. and Kettnaker V., 2000 "Discovery and Segmentation of Activities in Video," IEEE Transactions on Pattern Analysis and Machine Intelligent, Vol.22, pp.844-851.

Hu W., Tan T.,, Wang L., and Maybank S., 2004 "A Survey on Visual Surveillance of Object Motion and Behaviors" IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 34, No. 3.

Johansson G., 1973 "Visual Perception of Biological Motion and Model of its Analysis", Perception and Psychophysics, Vol.14, pp.201-211.

Kang H., Lee C.-W., and Jung K., 2004 "Recognition-Based Gesture Spotting in Video Games", Pattern Recognition Letters, Vol.25, No. 15, pp. 1701-1714.

Marchand-Maillet S., 2000 "Content-based *Video* Retrieval: An overview", Technical Report Vision.

Meier U., Stiefelhagen R., Yang J., and Waibel A., 2000 "Toward unrestricted lip reading," International Journal on Pattern Recognition and Artificial Intelligent, Vol.14, no.5, pp.571-585.

Oliver N. M., Rosario B., and Pentland A. P., 2000 "A Bayesian Computer Vision System for Modeling Human Interactions," IEEE Transactions on Pattern Analysis and Machine Intelligent, Vol.22, pp.831-843.

Owens J. and Hunter A., 2000 "Application of the Self-Organizing Map to Trajectory Classification," Proceedings of IEEE International Workshop on Visual Surveillance, pp.77–83.

Rumelhart D., Ortony A., 1964 "The Representation of Knowledge in Memory" In R.C. Anderson, R.J. Spiro, W.E. Montague (Eds.) Schooling and the acquisition of knowledge pp.99-135. Hillsdale, NJ: Erlbaum.

Rumelhart D.E., Hinton G.E., and Williams R.J., 1986 "Learning Representations by Back-propagation of Errors", Nature, Vol.323, pp.533-536.

Runeson S., 1994 "Perception of Biological Motion: the KSD-Principle and the Implications of a Distal Versus Proximal Approach". In G. Jansson, W. Epstein & S. S. Bergström (Eds.), Perceiving events and objects, pp.383-405.

Sumpter N. and Bulpitt A., 2000 "Learning Spatio-Temporal Patterns for Predicting Object Behaviour," Image and Vision Computing, Vol.18, No.9, pp.697–704.

Venkatesh Babu R. and Ramakrishnan K.R., 2004 "Recognition of Human Actions using Motion History Information Extracted from the Compressed Video", Image and Vision Computing, Vol.22, No.8, pp.597-607.

Viviani P. and Schneider R., 1991 "A Developmental Study of the Relationship Between Geometry and Kinematics in Drawing Movements" Journal of Experimental Psychology: Human Perception and Performance, Vol.17, pp.198-298.

Yoon H., Soh J., Bae Y. J., Yang H.S., 2001 "Hand Gesture Recognition Using Combined Features of Location, Angle, and Velocity", Pattern Recognition, Vol.34, pp.1491-1501.