An Image Retrieval Based Solution for Correspondence Problem in Binocular Vision

Alberto Amato*, Vincenzo Piuri
DTI - Università degli Studi di Milano
Via Bramante 65, 26013 Crema (CR)
Italy
{alberto.amato, vincenzo.piuri}@unimi.it

Vincenzo Di Lecce
DIASS - Politecnico di Bari
Via Alcide De Gasperi, 74100 Taranto (TA)
Italy
v.dilecce@aeflab.net

***Abstract*— Aim of this paper is to propose a solution to the correspondence problem in multi-camera systems. In these systems, two or more cameras are used to record the same scene from different view points. In this way it is possible to face the problem of occlusions in crowding scenes.**

In this work an object level motion detection algorithm is used and it is applied to the videos sampled by two cameras. The proposed approach does not require a calibration stage and it does not introduce any constraints about the camera positions. Once that the moving objects are detected, they are characterized using image retrieval techniques.

The system was tested using two cameras. Object detection and tracking are primary tasks in automatic video streaming analysis. The obtained results in terms of correct classifications rate seem to be encouraging because they highlight the ability of the system to work also in presence of crowding scenes.

Keywords; correspondence problem; visual feature based method, binocular vision, video surveillance system.

I. INTRODUCTION

Due to the rapid growth of the ICT, nowadays there is a significant spread of digital devices able to capture good quality videos. This fact allows the design and implementation of video acquisition systems using many cameras and often used in video surveillance applications. One of the biggest challenge in this research field is to develop a system that is able to automatically understand the sampled videos and rise an alarm in real-time when a danger situation is detected. In order to face this problem, in literature there are various studies about automatic video streaming analysis [1, 2, 3, 4]. These systems propose different approaches to semantic video analysis but they share the characteristic of working into the narrow domain. For example the method proposed in [4] works for football match videos while that proposed in [3] works for traffic monitoring, etc.

Another characteristic shared among these systems is the fact that they work analyzing the trajectories of some relevant points (typically the barycentres) representing the whole moving objects. For example, in [5] a system able to recognize 47 actions performed by different individuals is proposed. It analyzes the position of the hand barycentres.

All these systems have good performance but they fail in discovering motion when the view direction is parallel to the plane where the action is being performed. In other words, since they use a single camera based viewpoint, they have a bi-dimensional view of the world loosing the depth of field.

In order to face this problem, in literature the binocular and multi-camera approaches have been proposed [6, 7, 8]. These systems work using two (or more) cameras viewing the same scene from different view points. Using this approach it is possible to appreciate the depth of field solving the previous problem. Furthermore, it is possible to solve or at least reduce the problem of occlusions in crowding scenes.

On the other hand, this approach introduces the correspondence problem namely given a pixel or an object into the image sampled by one camera, where is it in the frame sampled by the other camera?

In literature various authors propose solutions to this problem introducing some constraints. For example, a widely applied constraint is that the acquisition system produces stereo images [9, 10]. In this case, the images are taken by two cameras with parallel optic axes and displaced perpendicular to the axes.

Using stereo pairs of images it is possible to assume that: stereo pairs are epipolar and the epipolar lines are horizontally aligned, i.e., the correspondence points in the two images lie along the same scan lines; the objects have *continuity* in depth; there is a one-to-one mapping of an image element between the two images (*uniqueness*); and there is an ordering of the matchable points [11].

From a geometric point of view, the corresponding problem can be successfully solved and the methods proposed in literature achieve excellent performance when applied to synthetic images. When these methods are applied to real world images, the main issues to be solved are: noise and illumination changes as a result of which the feature values for the corresponding points in the two images can differ; lack of unique match features in large regions; occlusions, and half occlusions. The wider used methods to solve this problem are: area based [10], feature based [12], bayesan network [13], neural networks [11, 14], etc.

The video surveillance systems using binocular (or multi-camera) vision use two different approaches to solve this

978-1-4244-8075-3/10/$26.00 ©2010 IEEE

problem: recognition based [15, 16] and geometry based [7, 8]. The latter methods give results more robust than the former but they have also a higher computational cost.

In this paper, the authors propose a recognition based method using techniques known in image retrieval field to solve the correspondence problem. This system must be seen as a stage of a longer processing chain aiming at semantic video analysis. The designed architecture can be easily implemented on parallel machines making the system suitable for time critical monitoring applications.

Comparing the proposed system with the others described in literature, its main advantages are: it does not require camera calibration and it has a low computational cost. The proposed approach uses a motion detection algorithm working at object level. Each detected object is characterized using two low level visual features: region-based color histogram and texture. The remaining part of this work is so organized: section II describes the proposed system, the experiments are shown in section III while conclusion and final remarks are reported in section IV.

II. PROPOSED SYSTEM

Nowadays, the systems for semantic video analysis follow a computational chain well defined:

Video acquisition → Frame extraction → Motion detection → Object detection → Object tracking → Semantic Analysis.

The proposed system uses two cameras installed in an arbitrary way. The only constraint is that the scenes recorded by the two cameras must have an overlapping zone. The greater is the overlapping zone, the greater is the area where the correspondence problem can be solved.

In this work, we assume that the models of the target objects and their motion are unknown, so as to achieve maximum application independence. In this condition, the most widely adopted approach for moving object detection with fixed camera is based on background subtraction [17]. The background is estimated using a model evolving frame by frame. The moving objects are detected by the difference between the current frame and the background model. A good background model should be as close as possible to the real background and it should be able to reflect as soon as possible the sudden change in the real background.

According to the taxonomy proposed in [18], the proposed system considers the following objects:

- Moving visual object (MVO): a set of connected pixels moving at a given speed
- Background: is the current model of the real background
- Ghost: a set of connected pixel detected as "in motion" by the subtraction algorithm but not corresponding to any real moving object.

The proposed system uses these three elements to implement an object oriented motion detection algorithm. It uses the knowledge about the segmented object to dynamically improve the background model. In order to classify the object after the blob segmentation the following rules are used:

1) <MVO> ← (foreground blob) ∧ (large area) ∧ (high speed)

2) <GHOST> ← (foreground blob) ∧ (large area) ¬ (high speed)

Once that a MVO is detected, it is described using an image retrieval technique. From each MVO two low level visual features (signatures) are extracted: region based color histogram and texture. These features were used in many Content Based Image Retrieval (CBIR) systems [19, 20, 21].

Color histograms have proved to be a powerful representation for the image data in a region. They have been used in several successful tracking systems which take advantage of their robustness in changing object pose and shape [22], [23]. The main drawback of this visual feature is that it discards all spatial information. This fact reduces specificity in the model, thus increasing the susceptibility of the tracker to distraction by the background or by other objects. In order to overcome this limitation, in this work, a region-based color histogram was used. The detected human silhouette is divided into three areas: head, body, legs. For each area, the color histogram is computed using the HSV (Hue, Saturation and Value) space color. The original color space of each MVO (RGB) is converted into HSV color space and then a histogram composed of 256 bins (16 hue, 4 saturation and 4 value levels) is computed for each region. This color space was preferred to the RGB color space because the former is closer than the latter to the human

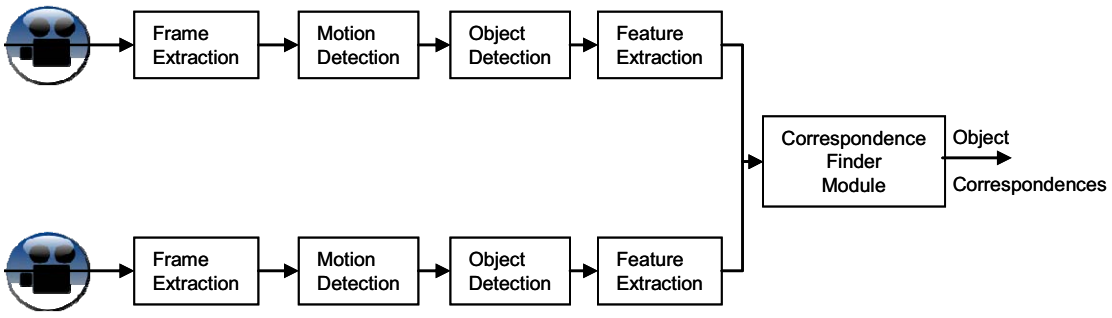

Figure 1 – Block diagram of the proposed system

color perception scheme [24].

Textures can be defined as 'homogeneous patterns or spatial arrangements of pixels that regional intensity or color alone does not sufficiently describe' [25]. Among contents based features, texture is a fundamental feature which provides significant information for image classification, for this reason it has been used in the proposed system. In order to describe the texture of each MVO, the approach proposed in [26] has been used. This approach is based on using Gabor filters and gives us a feature vector composed of 48 components.

Figure 1 describes a block diagram of the proposed approach. The block diagram shows the high level of parallelism of the proposed approach. Indeed, it is possible to design a parallel architecture where each machine processes the stream of a single camera (or of a group of cameras). The only sequential block is the last one namely the correspondence finder module. Nevertheless, this block can not be considered as a bottleneck of the architecture because from a computational point of view, it computes a not expansive task.

The main block of the proposed system are:

- Frame extraction: at a given time *t*, a frame is extracted by each camera. Let X_t and Y_t be the frames extracted at the instant *t* by T1 and T2 respectively (T1 and T2 are the two cameras as shown in figure 2). To this couple of frames are applied the following processes to solve the correspondence problem only for the detected moving objects. In this work, each second of the sampled video sequence is decomposed into 15 frames (on both cameras).

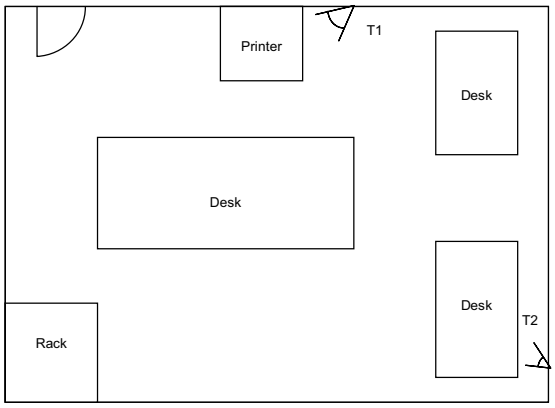

Figure 2 - A schematic view of the experiment setup. T1 and T2 are the two camera

- Motion detection: for each camera, the various frame are used to detect the moving areas according to the model above described.
- Object detection: for each frame, the moving areas are analyzed to detect MVOs according to the model above described. In particular, when the human body silhouette is composed of disjoint moving areas (due to various reasons such as noise, some parts of the body can look like the background, etc) simple geometric constraints are used to recover it. For each analyzed frame, the output of this module is the list of detected MVOs.
- Feature extraction: let O_{xt} and O_{yt} be the lists of detected MVOs in X_t and Y_t respectively. For each $x \in O_{xt}$ and y

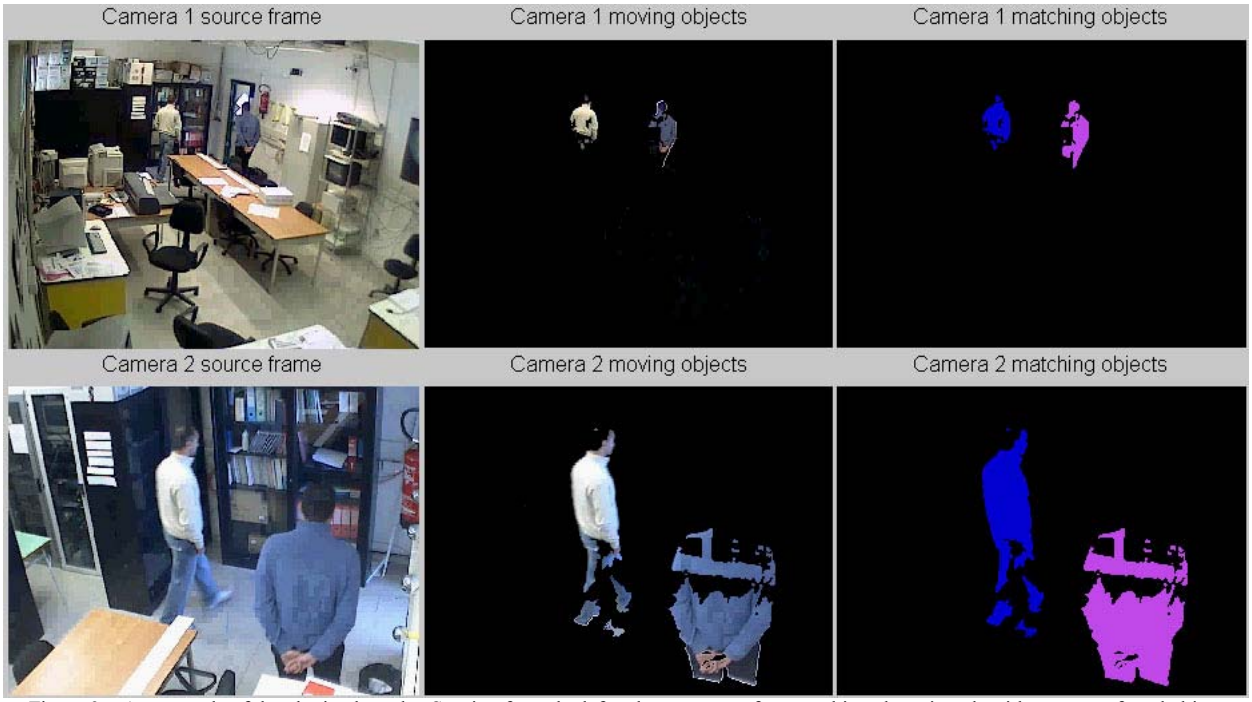

Figure 3 – An example of the obtained results. Starting from the left column: source frames, object detection algorithm output, found object correspondences. The corresponding moving objects into the two frames are colored using the same color.

$\in O_{yt}$, the region based color histogram and the texture features are extracted. Let $F(O_{xt})$ and $F(O_{yt})$ be the sets of visual features extracted by the lists of MVOs O_{xt} and O_{yt} respectively. $F(O_{xt})$ and $F(O_{yt})$ are sent to the "correspondence finder module".

- Correspondence finder: this module searches for the correspondences among the features in the sets $F(O_{xt})$ and $F(O_{yt})$. It finds the correspondences among the detected MVOs computing the distances among the elements of $F(O_{xt})$ and $F(O_{yt})$. In this work, a weighted normalized Euclidean distance is used. This kind of distance is used to consider the similarity contribute of both the features. In the proposed experiments, the weight is equal to 0.5 to give the same relevance to both the features.

III. EXPERIMENTS AND RESULTS

The system was tested carrying out various experiments in a room of our Faculty. Figure 2 shows the used experimental setup. T1 and T2 are the two used cameras. In these experiments two Axis 210 network cameras were used. These cameras are able to sample till 30 frames per second but in these experiments only 15 frames per second were used. This because a higher frame rate does not introduce more details in the correspondence problem process. Indeed, using more than 15 frames per second the differences between two successive frames sampled by the same camera are negligible. The cameras are installed at a distance of about three meters from the ground plane. The image in the upper left corner of figure 3 shows the view points of T2 while that in the bottom left corner shows the one of T1.

In order to test the system, various sequences were sampled where one or more people are moving in the room. Figure 3 shows an example of the obtained results. In the first column there are the frames as they are sampled by the cameras. T2 sees a cluttered environment. Indeed, the silhouettes of the walking people are partially occluded by two desks. The second column shows the output of the proposed object detection algorithm applied to both the cameras. The third column shows the results obtained by the proposed solution to the correspondence problem applied to these frames. The corresponding moving objects into the two frames are colored using the same color. Table I shows some statistics about the mean results obtained analyzing ten videos where various people are walking into the room. This table shows the obtained recall and the percentage of detected false matching for videos where there were respectively one person, two people and three people walking. The obtained results in terms of recall are almost constant among the three typology of experiments. This fact highlights the robustness of the proposed system to the partial occlusions. Indeed while the scenes shown in the first column of figure 3 have a constant partial occlusion for camera T2 due to the presence of the desks, during the experiments with more people, other dynamic partial (and often total) occlusions appear due to the relative motion of the walking people. The growing complexity of the scene is highlighted also by the growing percentage of detected false matches.

Figure 4 represents a situation where three people are walking in the room. This is one of the most critical example of dynamic occlusion that can occur with the proposed experimental setup. Indeed, T2 sees three people with an

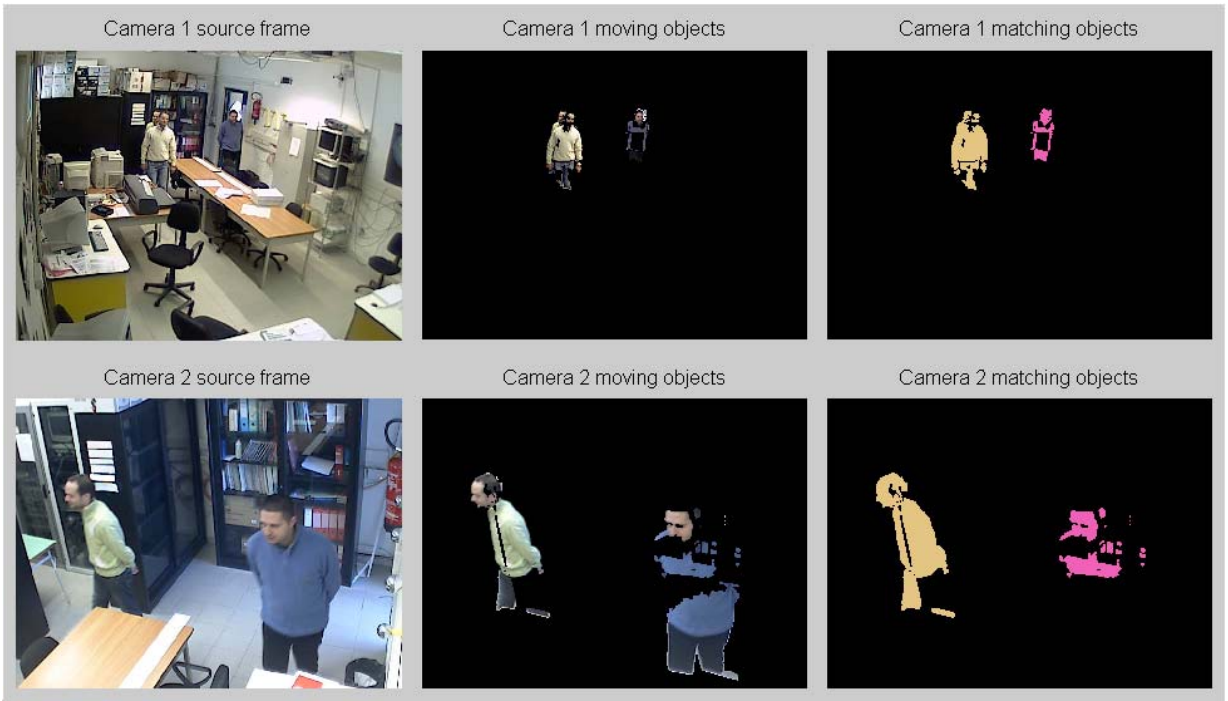

Figure 4 – An example of the obtained results. Starting from the left column: source frames, object detection algorithm output, found object correspondences. The corresponding moving objects into the two frames are colored using the same color.

Table 1 - results obtained analyzing ten different videos

Number of people	1	2	3
Recall (%)	73	76	74
% false match	5	23	25

almost total occlusion between two of them (more than 95% of one person is occluded by the other one). T1 sees only two people. Furthermore, it should be highlighted that the two self-occluding people are wearing dresses with very similar colors. The system is not able to discriminate between the two self-occluding people due to the almost total occlusion and to the strong similarity of their dresses. The result obtained by the correspondence finder module are shown in the last column. The two self-occluding people seen by T2 are considered as a single moving object and it is matched with the person in the bottom-left corner of the frame taken by T1. The other person seen by T2 are correctly matched with the person in the bottom-right corner of the frame taken by T1. The match between the two self-occluding people in T2 are matched with the "correct" person in T1. In other words, the person in T1 is one of the two very similar self-occluding people in T2.

It is difficult to compare the obtained performance with other works in literature. Indeed, in literature there are two main classes of works dealing with the correspondence problem:

1. works aiming to solve the problem for all the points of the images. These works try to obtain a dense disparity map and they evaluate the performance analyzing the percentage of correctly mapped pixels.
2. works using the multi-camera vision in video analysis applications (i.e. video surveillance systems). These works do not give too emphasis to the correspondence problem in the performance evaluation sections.

Starting from these considerations, the authors evaluated the obtained results in terms of recall and percentage of false matching because these parameters can give to the reader an objective idea of the system performance.

IV. CONCLUSIONS

In this work the authors propose a feature based solution to the correspondence problem in binocular vision systems.

The proposed system takes origin from the following considerations:

- Many systems for semantic video analysis work reducing the computation to the trajectory of some relevant points (typically the barycentre of the moving object). These systems suffer two problems: object occlusions and they fail in discovering motion when the view direction is parallel to the plane where the action is being performed;
- The binocular vision is a practicable approach to solve or at least reduce the problem of object occlusions and it is a valid solution to the depth of field losing problem in single camera systems;
- On the other hand, this approach introduces the correspondence problem. In literature there are various solutions to this problem but they often analyzes the whole images or introduce constraints about camera or ground plane calibration, etc.

The proposed approach tries to solve the correspondence problem for a small set of points reducing the computational cost. In particular, in this work this problem is solved only for the barycentre of each tracked object. Furthermore, this approach does not introduce any constraints about camera positions and calibration.

The first obtained results seem to be encouraging. Indeed they show that the system is able to work also in presence of partial occlusions. Indeed, due to the position of the two cameras, the silhouettes of the walking people taken by T2 are partially occluded by two desks. Furthermore, augmenting the number of walking people, the number of dynamic partial (or total) occlusions augments but the obtained results remain at the same level. This is due to the characteristics of the used visual features. Indeed, while the texture feature works on the whole MVO when the level of occlusion is low, the region based color histogram is able to work also when some parts of the MVO is occluded.

As figure 1 shows, the designed architecture can be implemented also on parallel machines. This fact makes the proposed system suitable also in time critical monitoring applications. Indeed, the proposed system should not be seen as a stand-alone system but it should be seen as a module of a more complex architecture aiming at semantic video analysis.

Future works will deal with the research of better visual features to describe the MVOs.

REFERENCES

[1] Yun Yuan; Zhenjiang Miao; Shaohai Hu; , "Real-Time Human Behavior Recognition in Intelligent Environment," Signal Processing, 2006 8th International Conference on , vol.3, no., 16-20 2006.

[2] Watanabe, K.; Izumi, K.; Kamohara, K.; Yamada, E.; , "Feature extractions for estimating human behaviors via a binocular vision head," Control, Automation and Systems, 2007. ICCAS '07. International Conference on , vol., no., pp.634-639, 17-20 Oct. 2007

[3] D Koller, J Weber, T Huang, J Malik, G Ogasawara, B Rao, and S Russel. Towards robust automatic traffic scene analysis in real-time. In Int. Conf. Pattern Recognition, pages 126--131, Jerusalem, Israel, October 1994

[4] Saito, H.; Inamoto, N.; Iwase, S.; , "Sports scene analysis and visualization from multiple-view video," Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on , vol.2, no., pp.1395-1398 Vol.2, 30-30 June 2004

[5] Cen Rao, Alper Yilmaz And Mubarak Shah, View-Invariant Representation and Recognition of Actions, International Journal of Computer Vision 50(2), pp. 203–226, 2002

[6] Hu, Z.; Uchimura, K.; , "U-V-disparity: an efficient algorithm for stereovision based scene analysis," Intelligent Vehicles Symposium, 2005. Proceedings. IEEE , vol., no., pp. 48- 54, 6-8 June 2005

[7] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axisbased correspondence between multiple cameras for people tracking," IEEE Trans. Pattern Anal. Machine Intell., vol. 28, no. 4, pp. 663–671, 2006.

[8] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," IEEE Trans. Pattern Anal. Machine Intell., vol. 25, pp. 1355–1360, 2003.

[9] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 7, pp. 675–684, Jul. 2000.

[10] S. Yoon, S. K. Park, S. Kang, and Y. K. Kwak, "Fast correlationbased stereo-matching with the reduction of systematic errors," Pattern Recognit. Lett., vol. 26, no. 14, pp. 2221–2231, Nov. 2005.

[11] Y. V. Venkatesh, S. Kumar Raja, and A. Jaya Kumar, "On the Application of a Modified Self-Organizing Neural Network to Estimate Stereo Disparity", in IEEE Transactions On Image Processing, Vol. 16, No. 11, November 2007, pp. 2822-2829.

[12] Lu Yang, Rongben Wang, Pingshu Ge, Fengping Cao, "Research on Area-Matching Algorithm Based on Feature-Matching Constraints", in 2009 Fifth International Conference on Natural Computation, pp. 208-213

[13] Jian Sun; Nan-Ning Zheng; Heung-Yeung Shum; , "Stereo matching using belief propagation," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.25, no.7, pp. 787- 800, July 2003

[14] Ruichek, Y.; , "A hierarchical neural stereo matching approach for real-time obstacle detection using linear cameras," Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE , vol.1, no., pp. 299- 304 vol.1, 12-15 Oct. 2003

[15] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easyliving," in IEEE International Workshop on Visual Surveillance, 2000, pp. 3–10.

[16] A. Mittal and L. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," in European Conference on Computer Vision, 2002, pp. 18–36.

[17] C. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real- Time Tracking of the Human Body," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780-785, July 1997.

[18] Rita Cucchiara, Massimo Piccardi, Andrea Prati, Detecting Moving Objects, Ghosts, and Shadows in Video Streams, in IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 25, no. 10, October 2003

[19] W.Pedrycz, A. Amato, V. Di Lecce, V. Piuri, Fuzzy Clustering with Partial Supervision in Organization and Classification of Digital Image, IEEE Trans. On Fuzzy System, Vol.16, no.4,pp. 1008-1026, August 2008. ISSN: 1063-6706

[20] A. Amato, V. Di Lecce, A knowledge based approach for a fast image retrieval system, Image and Vision Computing (2008), Volume 26 , Issue 11 (November 2008), pp. 1466-1480, ISSN:0262-8856

[21] Q. Iqbal, K. Aggarwal, Feature integration, multi-image queries and relevance feedback in image retrieval, in: Invited Paper, 6th International Conference on Visual Information Systems (VISUAL 2003), Miami, Florida, September 24–26, 2003, pp. 467–474

[22] S. Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms," Proc. CVPR, 1998, pp. 232-237.

[23] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 25, no. 5, May 2003, pp. 564-577.

[24] Ojala T, Rautiainen M, Matinmikko E & Aittola M, Semantic image retrieval with HSV correlograms, in Proc. 12th Scandinavian Conference on Image Analysis, Bergen, Norway, 621 - 627

[25] J. R. Smith, "Integrated Spatial Feature Image Systems: Retrieval, Analysis and Compression", Ph.D. thesis, Graduate School of Arts and Sciences, Columbia University, 1997

[26] Grgic, M.; Ghanbari, M.; Grgic, S.; Texture-based image retrieval in MPEG-7 multimedia system, in EUROCON'2001, Trends in Communications, International Conference on. Volume 2, 4-7 July 2001 Page(s):365 - 368 vol.2