Object-based video: extraction tools, evaluation metrics and applications

Andrea Cavallaro and Touradj Ebrahimi

Signal Processing Institute, Swiss Federal Institute of Technology (EPFL),
CH-1015 Lausanne, Switzerland

ABSTRACT

The representation of video information in terms of its content is at the foundation of many multimedia applications, such as broadcasting, content-based information retrieval, interactive video, remote surveillance and entertainment. In particular, object-based representation consists in decomposing the video content into a collection of meaningful objects. This approach offers a broad range of capabilities in terms of access, manipulation and interaction with the visual content. The basic difference when compared with pixel-based procedures is that instead of processing individual pixels, image objects are used in the representation. To exploit the benefits of object-based representation, multimedia applications need automatic techniques for extracting such objects from video data, a problem that still remains largely unsolved. In this paper, we first review the extraction techniques that enable the separation of foreground objects from the background. Their field of applicability and their limitations are discussed. Next, automatic tools for evaluating their performances are introduced. The major applications that benefit from an object-based approach are then analysed. Finally, we discuss some open research issues in object-based video.

Keywords: object-based video, segmentation, quality evaluation, video editing, video coding

1. INTRODUCTION

One of the goals of image analysis is to extract meaningful entities from visual data. A meaningful entity in an image or an image sequence corresponds to an object in the real world, such as a tree, a building, or a person. The ability to manipulate such entities in a video as if they were physical objects is a shift in the paradigm from pixel-based to content-based management of visual information.[1,2] In the old paradigm, a video sequence is characterized by a set of frames. In the new paradigm, the video sequence is composed of a set of meaningful entities. A wide variety of applications, ranging from video coding to video surveillance, and from virtual reality to video editing, benefit from this shift. The new paradigm allows us to increase the interaction capability between the user and the visual data. In the pixel based paradigm, only simple forms of interaction, such as fast forward and reverse, slow motion, are possible. The entity oriented paradigm allows the interaction at object level, by manipulating entities in a video as if they were physical objects. For example, it becomes possible to copy an object from one video into another.

The extraction of the meaningful entities is the core of the new paradigm. In the following, we will refer to such meaningful entities as *semantic video objects*. A semantic video object is a collection of image pixels that corresponds to the projection of a real object in successive image planes of a video sequence. The meaning, i.e. the *semantics*, may change according to the application. For example, in a building surveillance application, semantic video objects are people, whereas in a clothes shopping application, semantic video objects are the clothes of the person. Even this simple example shows that defining semantic video objects is a complex and sometimes delicate task. The process of identifying and tracking the collections of image pixels corresponding to meaningful entities is referred to as *semantic video object extraction*. The main requirement of this extraction process is *spatial accuracy*, that is, precise definition of the object boundary. The goal of the extraction process is to provide pixel-wise accuracy. Another basic requirement for semantic video object extraction is *temporal coherence*. Temporal

Further author information: (Send correspondence to A. Cavallaro)
A.Cavallaro: E-mail: andrea.cavallaro@epfl.ch, Telephone: +41 21 6934728
T. Ebrahimi: E-mail: touradj.ebrahimi@epfl.ch, Telephone: +41 21 6932606

coherence can be seen as the property of maintaining the spatial accuracy in time. This property allows us to adapt the extraction to the temporal evolution of the projection of the object in successive images. To maximize the benefits of the object-oriented paradigm, the semantic video objects need to be extracted in an automatic manner. To this end, a clear characterization of semantic video objects is required. Unfortunately, since semantic video objects are *human abstractions*, a unique definition does not exist. In addition, since semantic video objects cannot generally be characterized by simple homogeneity criteria (e.g., uniform color or uniform motion), their extraction is a difficult and sometimes loose task.

The paper is organized as follows. Section 2 reviews the different video object extraction techniques that can be used to extract video object from the input data. The methods to evaluate the performances of such techniques are presented in Section 3. Section 4 reports major applications that benefit from an object-based approach. Finally, Section 5 concludes the paper and provides a list of open research issues in object-based video.

2. EXTRACTION STRATEGIES

A semantic video object, as defined in the previous section, corresponds to a human abstraction. The extraction of semantic video objects from a video sequence requires that this abstraction be translated into rules. Once the rules are defined, they can be applied through an algorithm or a human operator. According to the amount of human intervention in the extraction process, we can classify the different approaches to semantic video object extraction in three classes: *manual*, *automatic* (unsupervised), and *semi-automatic* (supervised or interactive).

2.1. Manual extraction

In the case of manual extraction, the rules representing the semantic information are applied directly by the user. This procedure allows a perfect definition of the object boundaries. Spatial accuracy and temporal coherence are therefore guaranteed. However, this procedure is very time consuming, and obviously cannot be used for the large amounts of visual data available nowadays, for example, in content-based video retrieval. A manual approach is necessary in some cases, such as high quality film production ot the creation of a reference segmentation in order to assess the quality of automatic or semi-automatic extraction techniques.[3]

2.2. Automatic extraction

Fully automatic methods apply the rules defining the semantic objects in an algorithmic way. These rules are based on special characteristics of the scene or on specific knowledge (*a priori* information). They are derived for a specific application, or class of applications. A typical example of methods based on a specific set-up of the scene is the *blue screen* approach (a.k.a. chroma-keying). Examples of methods based on *a priori* information are template matching, face detection, and moving object segmentation.

2.2.1. Chroma-keying

In the *blue screen* approach to automatic extraction, objects in the real world are filmed in front of a uniformly colored background (usually blue or green). The background is then eliminated by discarding pixels with the known background color. The blue screen approach provides good spatial accuracy and temporal coherence. However, it is not generically applicable, because it requires a specific set-up of the scene. Besides, special care is required for the lighting of the scene, to avoid shadows.

2.2.2. Depth-keying

To overcome some limitations of chroma-keying, a depth-keying technique can be used. In the depth-keying approach, depth information is used to separate foreground objects from the background. A depth map is computed based on active or passive methods. *Active methods* couple traditional cameras with depth sensors. This approach provides good spatial accuracy in the segmentation results. However, objects can be segmented only when they are close to the camera, at a distance that depends on the range of the depth sensor. For this reason, active approaches are mainly used in indoor applications. *Passive methods* are based on the use of multiple cameras. A depth map is computed from the information coming from stereo cameras. One advantage of passive methods is that they can be applied on both indoor and outdoor scenes. On the other hand, they provide with low spatial accuracy in the segmentation results when objects have little texture.

2.2.3. *A priori* information

In this approach, some knowledge of the objects we want to extract substitutes the knowledge of the color of the background of the blue screen approach. Template matching, face detection, and moving object segmentation are typical examples of methods based on *a priori* information. If the shape of the object we want to segment is known *a priori*, *template matching* can be used to implement the semantics. In this case, the extraction method will look for specific object features in terms of geometry. If we want to segment faces of people, color-based segmentation can be used. The *face detection* task will consist in finding the pixels whose spectral characteristics lie in a specific region in the chromaticity diagram.[4] For extracting moving objects, *motion information* can be used as semantics. The motion of a moving object is usually different from the motion of background and other objects. For this reason, many extraction methods make use of motion information in video sequences to automatically extract semantic objects.

2.3. Semi-automatic extraction

Fully automatic extraction techniques are still in their infancy, because translating the properties of a semantic object into extraction criteria is a difficult task. For this reason, many semi-automatic strategies have been proposed as a trade-off between a fully automatic strategy and the manual extraction of video objects. The principle at the basis of semi-automatic techniques is the *interaction* of the user during some stages of the extraction process, where the semantic information is provided directly by the user. After the user provides the initial segmentation of the video object, a tracking mechanism follows its temporal evolution in the subsequent frames thus propagating the semantic information. Since tracking tends to introduce boundary errors, the object boundaries need to be modified and updated either via further interaction[5] or according to some low-level homogeneity criteria.[6] Errors are generally due to imperfections of contours or to the appearance of new objects in the scene. According to the choices that the user makes in the interaction, semi-automatic techniques can be classified as *feature-based*, *contour-based*, or *region-based*. These three approaches may be used either separately or in combination,[7] thus allowing good flexibility.

2.3.1. Feature-based interaction

In the case of feature based interaction, the user selects some pixels belonging to the object that exhibit characteristic color/texture properties. These pixels are used as basis for the extraction: they are characterized by their features and the remaining pixels are then classified accordingly.[8] The advantage of this method is that is not necessary to demarcate the objects precisely. However, there might be problems in the connectivity of the object. A further interaction step is required to overcome this problem.

2.3.2. Contour-based interaction

Instead of selecting a set of points belonging to the object, the user can mark its contour.[5] This is the principle of contour-based extraction methods. The contour may either be defined as a set of control points or as a sketch of the object. When only a few points are provided, they are automatically connected. The precision of the sketch is not critical. In the case of a rough sketch, an algorithm is required to adjust the boundaries to the real ones. These methods provide very precise object contours, but they are usually slower than feature-based techniques.

2.3.3. Region-based interaction

Image regions can finally be used to semi-automatically extract video objects. In this case, the user interacts with the result of a preliminary segmentation of the image into regions.[6] The user marks some of these regions as corresponding to a semantic object. These are then automatically merged to obtain the shape of the semantic video object.

User interaction provides a simple way of integrating semantics into the extraction process, and is more efficient than manual extraction, since usually limits human intervention to one frame only. However, one of the disadvantages of a semi-automatic approach is the inability to detect new objects, as well as the fact that it is time consuming because of the necessity of user intervention.

3. EVALUATION METRICS

Common practices for evaluating object extraction results are based on human intuition or judgment (*subjective evaluation*) and consist in *ad hoc* subjective assessment by a representative group of observers. A significant number of observers is required to produce statistically relevant results, thus making subjective evaluation a time-consuming and expensive process. To avoid systematic subjective evaluation, an automatic procedure is desired. This procedure is referred to as *objective evaluation*.

Quality metrics for objective evaluation of object segmentation may judge either object segmentation algorithms or object segmentation results. The metrics are referred to as analytical methods or empirical methods, respectively. *Analytical methods* find their origins in region segmentation and evaluate segmentation algorithms by considering their principles, their requirements and their complexity.[9] The advantage of these methods is that an evaluation is obtained without implementing the algorithms. However, because of the lack of a general theory for image segmentation, and because segmentation algorithms may be complex systems composed of several components, not all properties (and therefore strengths) of segmentation algorithms may be easily evaluated. *Empirical methods*, on the other hand, do not evaluate segmentation algorithms directly, but indirectly through their results. To choose a segmentation algorithm based on empirical evaluation, several algorithms are applied on a set of test data that are relevant to a given application. The algorithm producing the best results is then selected for use in that application. Empirical methods can be classified into two groups, namely *goodness methods* and *discrepancy methods*.

3.1. Empirical goodness methods

Empirical goodness methods evaluate the performance of an algorithm by judging the quality of a segmented image. The evaluation is based on measuring the desirable properties of a segmentation. Here these properties are referred to as *goodness parameters*. Goodness parameters represent the properties of the *ideal* object segmentation. Defining the ideal object segmentation is a difficult task. Therefore goodness parameters are usually established according to human intuition. In general, desirable properties for a partition pertain to the interior and the shape of each segment, and to comparisons with adjacent segments: boundaries of each segment should be simple, not ragged, and should be spatially accurate; adjacent segments should have significantly different values with respect to their uniformity characteristics. In[10] the goodness parameters are based on the analysis of the segmentation boundaries and rely on color and motion information. The differences of color and motion parameters along the boundary of the estimated video objects are computed. An intra-frame and inter-frame histogram difference along the estimated boundary is considered for color. Similarly, different motion values are expected on the two sides of the object boundary. This metric is based on fairly restrictive assumptions, for example, that the color histogram is stationary from frame to frame, and that the color histogram of the background is different from that of the object.

The main advantage of methods based on goodness parameters is that they do not need a reference segmentation. However, evaluating the goodness of a semantic partition is not a well-defined task, because the semantic partition corresponds to a human abstraction, the semantics can change, and the semantics cannot be easily quantified. Finally, ad hoc rules need to be defined according to the application.

3.2. Empirical discrepancy methods

Empirical discrepancy methods also evaluate a object segmentation algorithm based on the quality of its results. Here the object segmentation result is compared to a reference segmentation. This reference segmentation represents the *ground truth*, or the ideal segmentation, and can be generated either manually or via a reliable procedure. The result of the evaluation is a disparity between the reference segmentation and the actual segmentation result. Discrepancy methods are based on directly measuring the deviation between two partitions. The deviation is evaluated through *discrepancy parameters*, which characterize each method.

Discrepancy parameters are based on the spatial and temporal deviations. These deviations may be appropriately weighted to take visually desirable properties of a segmentation mask into account. For example, pixel errors are separated into two groups, those that belong to the result but not to the reference (false positive) and those that belong to the reference but not to the result (false negative).[3] Furthermore the temporal stability of the segmentation mask shape may be considered. The discrepancy parameters may also be formulated as

misclassification penalties regarding shape and motion errors.[11] The discrepancy parameter spatial accuracy is defined in[12] by shape fidelity, geometrical similarity, edge content similarity, and statistical data similarity. Discrepancy parameters are combined over the time interval, to assess the quality of the entire spatio-temporal segmentation. The parameters are weighted so as to combine them in correct proportions and to match evaluation results produced by human viewers.

In applications where the final judge of quality is a human being, it is also important to consider the human visual system to design a quality evaluation procedure, in addition to the objective discrepancy parameters. Traditional evaluation methods do not consider this aspect and just consider objective criteria, such as discrepancy between two results. One distinctive feature of the method in[12] is the evaluation of object relevance for judging the quality of segmentation. The overall segmentation quality depends on the estimated importance of segmented objects in the scene.

4. APPLICATIONS

Object-based video can support a large variety of content-based applications, ranging from video analysis to video coding, from video manipulation to interactive environments. These applications can be schematized by the block diagram presented in Figure 1. Furthermore, application-dependent feedback can be used to improve the performance of semantic video object extraction by exploiting application specific information. An overview of the above-mentioned content-based applications, as well as examples of specific implementations are given in this section.

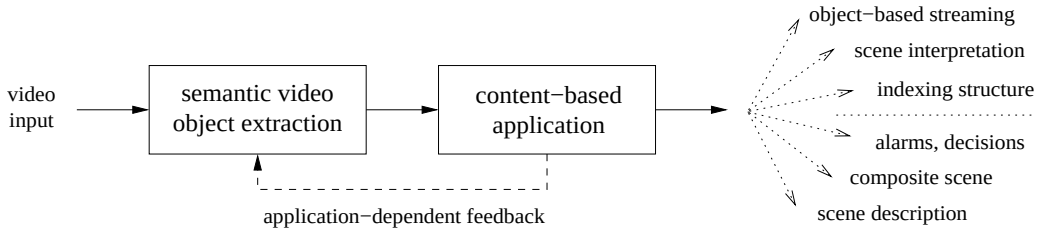

Figure 1. Object-based video enables a wide range of applications, such as object-based video coding, computer vision, scene understanding, and content-based indexing and retrieval. Application specific information can be used to improve the extraction process through interaction with the extraction module

4.1. Video editing and interactivity

The extraction of semantic video objects opens new dimensions in *interactions with video content*: the user can manipulate video objects as they were physical objects. Therefore the object-based approach facilitates creativity, by allowing the user to put together video objects from different sources, to produce new content. This feature has been obtained in the past with the manual techniques described in Section 2.1, or through bluescreens and chroma keying (Section 2.2). These techniques are time consuming and expensive, and therefore only a limited amount of video material can be manipulated manually. Automatic and interactive video object extraction is a component technology that can reduce manual interaction, and also cut production costs. In this framework, video object extraction can be seen as intermediate digital technology that represents a natural extension to cinematography. Once the video objects are extracted, new and richer content is created through editing, hyperlinking, special effects, and combination with computer-generated objects.

The creation of new content via video editing may also be complemented by associating more information with each video object. In the same way as hyperlinks are used on the world wide web when accessing text documents and pictures, hyperlinks can be associated to semantic video objects. This functionality is known as *hypervideo*.[13] Static and moving objects in a video can be tracked and made sensitive with a link at any time in the stream, to enhance the interactive experience. To implement this functionality, the semantic partition is used as a mask for defining the hyperlinks related to objects. This feature is particularly interesting for e-commerce

applications. While watching a football match, for example, the user is allowed to select a player (e.g., with the mouse) and get information about his biography, statistics about his performance in the current match and in previous ones. Furthermore, by selecting a part of the object like the t-shirt or the shoes, the user may get information about the sponsor, and the price of the sportswear.

The video editing features described so far are used to compose new content by assembling information, such as shots and video objects, from different sources. In addition, these sources can be modified by external elements that are not related to the characters in the scene. This process is known as *effect animation*. Effect animation allows the user to manipulate the video content in order to create special effects, such as object deformation, object freeze, stroboscopic effect, and artificial video object trajectory modification. *Object deformation* is made possible by changing the shape descriptors of an object. This modification is then followed by texture mapping, to adapt the texture to the new shape. *Object freeze* is the process of stopping the action of one object by simply repeating its position for a duration of time. The *stroboscopic effect* consists in repeating object freeze for successive time instants in a certain time interval. This effect is particularly useful in sports applications, to visualize the movements of a sports person. This allows one to compare the behaviour of different athletes and consequently correct possible errors. *Artificial video object trajectory modification* is achieved by modifying the trajectory descriptor of an object. In addition to this, the speed of an object may be modified by interpolating, or by skipping some positions.

4.2. Augmented reality and intelligent environment

Augmented reality is an emerging technology by which a user's view of the real world (real environment) is augmented with additional information (virtual objects) from a computer model.[14] Here, video object extraction serves to extract objects, that are then inserted in a virtual background. This offers multimedia authors the capability of designing immersive and interactive narratives that involve real people into a universe of pictures, graphics and associated designs created by the author. One of the goals of augmented reality is to create narrative spaces and interactive stories that mix graphical elements with the live input of several cameras. Subjects in front of these cameras get themselves immersed within the visual ambiance and they are therefore involved within the narrative, which they are able to interact with, through their behavior. Furthermore, other people may be viewing the mixed images on other screens, and may even be able to interact with the system themselves. In the above scenarios two types of objects may be considered: artificial objects, which have been artificially created by an artist, and real objects, which exist in the real world and are extracted from a real scene by separation from the background. In this framework, users participate in a common scenario where artifical and real objects coexist. This permits not only the enhancement of narrative spaces, but also the creation of telecommunication environments which provide an "ideal virtual space with sufficient reality essential for communication".[15] Such environments are referred to as *immersive enviroments*. They can be used in immersive games and interactive or intelligent environments.

4.3. Video coding

The decomposition of the scene into meaningful objects can improve the coding performance over low-bandwidth channels. *Object-based video compression* schemes, such as MPEG-4,[16] compress each object in the scene separately. For example, it is useful for wireless applications where limitation in available bandwidth require optimization of video coding. The compression rate of each object may be a function of its importance in the scene. Thus the overall bit rate can be lowered while preserving the perceived image quality.

4.4. Video indexing

Quantitative descriptors may also be generated from each semantic video object. The description of visual content may span different abstraction levels. It can describe the low level (perceptual) features of the content. These include features such as color, texture, shape and motion. At the other end (e.g., high level of abstraction), it can describe conceptual information of the real-world being captured by the content. Intermediate levels of description can provide models that link low-level features to semantic concepts. In addition, because of the importance of the temporal nature of multimedia and sensitivity of multimedia concepts to context, dynamic aspects of content description need also to be considered. These aspects have been addressed by the MPEG-7

standard,[17] officially referred to as *Multimedia content description interface*, where the primary objective is to facilitate the description, identification and access of audio-visual data. The representation of audio visual information in a content-based manner allows for simple to sophisticated description of the content of audiovisual information. This enables applications such as searching and filtering where specific content is of interest. Individual video objects may be stored and accessed, instead of just video clips or images, when a video object and its corresponding indexing structure are stored in a database. The use of such descriptors is useful not only for indexing but also for other applications such as video summarization, universal multimedia access, and computational visual surveillance. In the framework of video summarization, the image representation obtained by segmenting the moving objects is used to extract indexing criteria for summarizing a scene in a few structured semantic descriptors. These descriptors make it possible to create a *scalable* summary for universal multimedia access. Such a summary allows the user to browse the video database at several levels of detail. It also supports better indexing and organization of the description structures, to improve the efficiency and effectiveness of searching, filtering, and description manipulation.

4.5. Security

The evolution of *video surveillance systems* has led from the so called first generation CCTV systems to the second generation PC based systems.[18] This has favored the introduction of automatic digital image processing techniques to assist a human operator in video surveillance tasks, thus reducing the need for manual continuous monitoring. Human intervention can therefore be limited to higher level tasks. The user interprets only critical situations, makes decisions in ambigous cases, and chooses the most appropriate action when an automatic alarm is raised. Semantic modeling may help computational visual surveillance in several ways. In simple cases, the results of the change detection (semantic segmentation) are good enough, such as for intrusion detection. More complex systems require tracking of video objects, to achieve high level understanding.[19]

5. CONCLUSIONS

In this paper we presented an overview of the actual tools available for extracting video objects. Techniques for evaluating the performances of such tools have been reviewed. Finally, we discussed applications that benefit from the development of reliable and efficient object-based algorithms.

Despite a growing interest in this domain, object-based video is still at its infancy. Automatic object extraction tools are often very complex algorithms requiring fine tuning. These tools depend on the specific application and their performances heavily depend on illumination conditions. Many interesting research directions remain to be explored in this domain. For example, bridging the gap between low-level features and semantic description is a key factor for many applications. What can be consider low-level? What can be consider high-level? What interactions can be envisaged between low-level an high-level? What is the relationship between computer vision and artificial intelligence in the framework of object-based video?

Furthermore, there is still not a generally accepted definition of object segmentation. Different applications require different outputs from an object extraction tool. These outputs can be binary masks, binary masks enriched with texture, several layers, or 3D descriptions. A critical step would be to define a standard for object-based video which goes beyond the specifications provided by MPEG-4. In addition to this, multimodality could facilitate more robust and automatic segmentation and tracking of video objects under diverse conditions, such as multiple non-rigid objects undergoing self and mutual occlusions. Finally, evaluation procedures are not mature yet. Metrics working with and without ground-truth need to undergo further development in order to enable progress in object-based video.

REFERENCES

1. M. Kunt, A. Ikonomopoulos, and M. Kocher, "Second generation image coding techniques," *Proceedings of the IEEE*, vol. 73, no. 4, pp. 549–575, 1985.
2. H. G. Musmann, M. Hötter, and J. Ostermann, "Object-oriented analysis-synthesis coding of moving images," *Signal Processing: Image Communication*, vol. 1, no. 2, pp. 117–138, 1989.

3. X. Marichal and P.Villegas, "Objective evaluation of segmentation masks in video sequences," in *Proceedings of X European Signal Processing Conference (EUSIPCO)*, Tampere, Finland, pp. 2193–2196, 2000.
4. L. Gu and D. Bone, "Skin colour region detection in MPEG video sequences," in *Proc. of 10th International Conference on Image Analysis and Processing*, Venice, Italy, pp. 898–903, 1999.
5. C. Gu and M.-C. Lee, "Semiautomatic segmentation and tracking of semantic video objects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 572–584, 1998.
6. R. Castagno, T. Ebrahimi, and M. Kunt, "Video segmentation based on multiple features for interactive multimedia applications," *IEEE Transactions on Circuits and System for Video Technology*, vol. 8, pp. 562–571, September 1998.
7. B. Marcotegui, F. Zanoguera, P. Correia, R. Rosa, F. Marques, R. Mech, and M. Wollborn, "A video object generation tool allowing friendly user interaction," in *Proceedings of International Conference on Image Processing*, pp. 391–395, 1999.
8. E. Chalom and V. Bove, "Segmentation of an image sequence using multi-dimensional image attributes," in *Proceedings of International Conference on Image Processing*, pp. 525–528, 1996.
9. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, pp. 1335–1346, 1996.
10. C. Erdem, A. M. Tekalp, and B. Sankur, "Metrics for performance evaluation of video object segmentation and tracking without ground-truth," in *Proc. Int. Conference on Image Processing*, Thessaloniki, Greece, 2001.
11. C. Erdem and B. Sankur, "Performance evaluation metrics for object-based video segmentation," in *Proc. X European Signal Processing Conference*, vol. 2, (Tampere, Finland), pp. 917–920, September 2000.
12. P. Correia and F. Pereira, "Objective evaluation of relative segmentation quality," in *Proc. Int. Conference on Image Processing*, vol. 2, (Vancouver, Canada), pp. 308–311, September 2000.
13. H. Jiang and A. Elmagarmid, "Spatial and temporal content-based access to hypervideo databases," *International Journal on Very Large Data Bases*, vol. 7, no. 4, pp. 226–238, 1998.
14. W. Mackay, "Augmented reality: linking real and virtual worlds," in *Proceedings of Advanced Visual Interfaces (AVI)*, L'Aquila, Italy, pp. 1–9, 1998.
15. P. Milgram and K. F., "A taxonomy of mixed reality visual displays," *IEICE Transactions on Information Systems*, vol. E77–D, no. 12, pp. 52–62, 1994.
16. T. Ebrahimi, "MPEG-4 video verification model: A video encoding/decoding algorithm based on content representation," *Signal Processing: Image Communication*, vol. 9, pp. 367–384, 1997.
17. T. Sikora, "The MPEG-7 visual standard for content description – an overview," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 696–702, 2001.
18. C. Sacchi and C. Regazzoni, "Multimedia communication tecnhiques for remote cable-based video-surveillance systems," in *Proc. of 10th International Conference on Image Analysis and Processing (ICIAP)*, Venice, Italy, pp. 1100–1103, 1999.
19. A. Cavallaro, D. Douxchamps, T. Ebrahimi, and B.Macq, "Segmenting moving objects: the MODEST video object kernel," in *Proceedings of Workshop on Image Analysis For Multimedia Interactive Services (WIAMIS)*, Tampere, Finland, 2001.