Proceedings of the International Conference on Machine Vision and Machine Learning
Prague, Czech Republic, August 14-15, 2014
Paper No. 66

Depth Based Context Modeling and Classification in Video-surveillance

Nour Charara[2], Omar Abou Khaled, Elena Mugellini
University of Applied Sciences of Western Switzerland
CH-1705 Fribourg, Switzerland
[2]EDST, Lebanese University
Beirut, Lebanon
nour.charara@edu.hefr.ch; omar.khaled@edu.hefr.ch; elena.mugellini@edu.hefr.ch

Iman Jarkass
University Institute of Technology, Lebanese University
Saida, Lebanon
ijarkass@ul.edu.lb

Maria Sokhn
Business Information Systems,
HES-SO/Wallis
CH-3960 Sierre, Switzerland
maria.sokhn@hevs.ch

Abstract - With a dedicated definition of 'Context' in image understanding systems, we present in this paper a novel context modelling and classification system. The main goal behind multimodal context modelling is to identify the context type from video-surveillance footage of multipurpose halls. First, the distribution of the different zones in a multipurpose hall is automatically captured using a dedicated depth based segmentation method. The discriminative description is illustrated by extracting five semantic features according to depth zones. These features are processed with the Transferable Belief Model to propose a classification. Results show the validity of the method for context recognition.

***Keywords*:** Context modelling, Pattern recognition, Scene segmentation, Video-surveillance.

1. Introduction

A lot of useful information that doesn't appear in the scene or in the still image helps the observer in understanding the scene contents. Since the context exploration is an indisputable aspect of human visual perception, there is a deal of evidence that context facilitates object detection and event recognition in computer vision systems. The context can assist to disambiguate degraded images/ sequences of images that cannot be recognized in isolation. Moreover, context provides critical information that can serve: to supplement the available information from the scene itself, to enforce the assumptions of the understanding solutions, and to access relevant nonlocal information.

A multipurpose hall (Figure. 1) offers many possible usages and it is designed to fit various kinds of events, such as a movie theatre (cinema), conference hall, or even for ceremonies. This variety of uses also implies a variety of contexts and consequently a large number of possible events. Despite the large number of machine vision research works that exploited the contextual information to improve the recognition task, considering the 'Context' as central issue is still rarely treated in the literature. The

context modeling in this work defines formalism for a generic context description. This description aims to identify the active context type in the video-surveillance footage of multipurpose halls.

In video understanding domain, many formal definitions of *Context* are proposed, trying to answer to an essential problematic: how to define the context in order to model it. In natural images, objects are strongly expected to fit into a certain relationship with the scene, and context gives access to that relationship. (Marques et al., 2010) expand the definition provided by (Wolf and Bileschi, 2006)and call context "any information that might be relevant to object detection, categorization and classification tasks, but not directly due to the physical appearance of the object". An interesting issue is raised by (Dourish,2003). He tries to explore from the previous definitions the notion of context in human-computer interactive process. So he inferred some assumptions and commitments: a) context is a form of information, b) context is stable, c) context and activity are separable and activity happens within a context. Several approaches have been used in previous work to model context. In almost works, the spatial information is a common type, where the spatial relationship between object/event of interest with the surrounding environment is the most relevant contextual information. From this point, the visualrepresentation of context using a scene map as a support is detailed in [Bremond and Thonnat 1988];a 2D map of polygonal zones with semantic description (entrance and exit zones)is used.A priori, these zones are manually defined offline by a human operator.Such segmentation and spatial reasoning improve the computation of different video understanding tasks. Other approaches detach themselves from the manual segmentation or context annotation and propose using graphical descriptors for automatic context representation. An interesting technique called *spatial envelope* [Oliva and Torralba 2001] is proposed. It acts as a global descriptor of the context and is capable to capture the gist of a scene. This technique is based on a set of perceptual dimensions, which define the dominant spatial structure of the scene: naturalness, openness, roughness...etc. Oliva and Torralba have shown that the spatial envelope can distinguish between 8 different categories of scene contexts (city, street, highway…etc).

Fig. 1.Typical multipurpose hall scene.

The main motivation behind our method is the key difference from all state-of-the-art approaches; the need for an adaptive model for different contexts in the same place (the multipurpose hall). Hence the particularity, a spatial modelling is not sufficient to gather all the discriminative features for the different context types. This paper is organized as follows. The next section introduces the framework DBCoM (Depth Based Context Modeling) and gives an overview of the system and its concept. Section 3 and 4 detail the DBCoM modules including the semantic features extraction and the TBM based classification part. Section 7 presents the experimental results illustrating the performance of the proposed method. Finally, the last section concludes the paper and presents the future work.

2. DBCoM Concept & Architecture

As the definition of the context depends on the process nature, a more specific definition to this project scope is required. So, a generic context definition is proposed*: Context is the environment where an event can occur. The isolation of this event, from its environment, leads to another interpretation of this event.*This definition has guided -to some degree- the way this method was proposed. This definition is oriented to cope with the interpretation process in image understanding systems.

It is hard (or even impossible) to model the entirety of all elements and actions in a scene. This approach is based on taking consideration the spatial distribution of the multipurpose hall with respect to

the depth, in other word the analysis process follows the apportionment of the depth zones, which are determined automatically from the filmed scene. The two key assumptions underlying depth based modeling are: 1) the image of the multipurpose hall is taken so as to cover most of the hall components in a single frame, 2) The hall zones are distributed longitudinally (as the traditional manner). These assumptions are not very strong compared to the ones generally assumed (Divvala 2009) such as multi-view point and synchronized camera network.

We argue this segmentation based strategy that in our used case (Multipurpose hall); we can clearly observe that it can be segmented into three zones, from the camera's focal plane; first the spectator seats area, then, more deeply, the sidewalk and last, the theater or the onstage or the wide screen (according to the usage). Each zone has a set of special features that can be essential in context classification decision. Some features intertwined at context/zone level, e.g. some features extracted from a zone x that characterizes a context A, can reflect a context B when they are extracted from another zone y. Consequently, the features are extracted from each zone separately.

The depth based context modeling is a way to describe the scene context using semantic properties. These properties provide a discriminative description for the different context candidates without taking in consideration the local objects. Conceptually, the high-level information or semantic features that are used in our system is presented in section 3. After extracting different semantic features from the scene according to the depth zones, these information sources are processed and combined within the transferable belief model (TBM) in order to propose a classification. Fig. 2 outlines the global view of the system, it consists of four stages: vanishing point detection, dynamic extended rectangle based scene segmentation, semantic features extraction and information fusion.

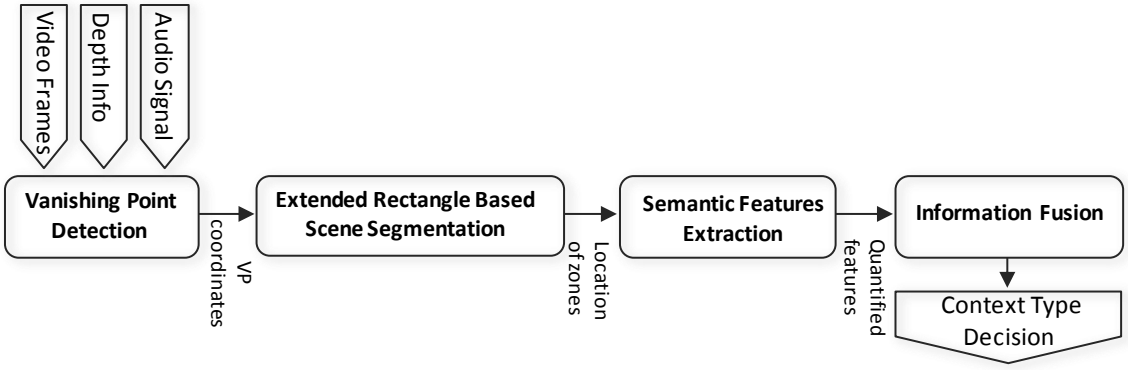

Fig. 2. DBCoM architecture overview.

The system takes three types of input from different sources: (1) **Depth Information**: the depth map of the scene is performed through a passive depth computation method using a stereo vision system. Two calibrated cameras with known physical relation were correlated. The correspondences of common pixel values (from left and right image) were found. By triangulation the distance between the correspondence areas was calculated. (2) **Audio signal**: a distributed network of microphones is equipped to extract low level audio parameters. These parameters are used to classify the sound segments and to locate the sound source. (3) **Video Frames**: the visual information in this system is exploited by two ways. First, as a sequence of frames with low frame rate. Or, the visual information is extracted from a still image (single camera shot). Both types are captured from the same fixed video-surveillance camera.

The main advantages of this method can be presented after this description. First, this method decreases the extracted data redundancy, when extracting just the adequate features for each zone. Second, this scheduling spirit leads to reduce the analysis complexity, and also to increase the robustness of context classification decision.

2.1. Scene Segmentation

Basing on the dedicated segmentation strategy DERSS (Dynamic Extended-Rectangle Scene Segmentation) (Charara 2013), the distribution of the different zones in a multipurpose hall is automatically captured without intervention from a human operator or a priori information. As well as the visual cues exploration (the texture), this method is also based on depth information to increase reliability in the comprehension of the scene elements.

3. Semantic Features

3.1. Audio Source Location ASL

This semantic feature aims to identify the zone(s) that is/are the origin of the audio source. Using three omnidirectional microphones collected in a microphone array, the TDoA (Time Difference of Arrival) method is applied to compute the coordinates of the audio source. The cross-correlation is the basic method that estimates the time delay between the three received signals. Indeed, the distances between the microphones and the sound source are given by the formula

$$d(M_i\,;S) = \sqrt{(x_i - x)^2 + (y_i - y)^2} \tag{1}$$

with x, y are the audio source$S$ coordinates and x_i, y_i $(i \in \{1,2,3\})$are the microphones coordinates.

The difference of distance δ_{12} between $d(M_1\,;S)$ and $d(M_2\,;S)$ is given by

$$\sqrt{(x_1 - x)^2 + (y_1 - y)^2} - \sqrt{(x_2 - x)^2 + (y_2 - y)^2} = \delta_{12} \tag{2}$$

with $\delta_{12} = c \cdot \tau_{12}$, where c is the sound celerity, hence the parabola equation:

$$\frac{\sqrt{(x_1\text{-}x)^2+(y_1\text{-}y)^2}-\sqrt{(x_2\text{-}x)^2+(y_2\text{-}y)^2}}{c} = \tau_{12} \tag{3}$$

Similarly, by rotating the indices, we obtain the equations of $\tau_{13}$and τ_{23}. The source coordinates are obtained by determining the intersection of these hyperbolas.

3.2. Screen Change Rate (SCR)

We benefit from the immobility of the equipped camera to crop the screen area (zone1) from all the video frames. The cropped frames will be analysed by the screen change detector to obtain the rate of change using edge detection based method. Our electronic screen change detection method is generally employed to detect major changes in video streams and to overcome the small changes sensitivity problem. In order to encounter the multiplicity challenge that rises from the multipurpose capacity, and to overcome problems concerning time-varying illumination and the small moving object problems, our proposed method first submits each video frame to Canny edge extractor. Then, the system finds features characterizing a scene/slide change detection using a technique inspired from the MPEG-7 Edge Histogram Descriptor. In particular, the cropped frame is segmented into 4×4 sub-images. Each sub-image is represented by a 5-bins edge distribution histogram where each bin survey one type of the five directional edge types: horizontal, vertical, diagonal-1 (+45º), diagonal-2 (-45º) and non-directional edge. A scene/ slide change detection between two successive frames f_i and $f_{i+1}$is inferred when the dissimilarity measure (eq. 4) exceeds a specific threshold.

$$d(f_i, f_{i+1}) = \sum_{j=1}^{j=\mathrm{N}} \sum_{k=1}^{k=K} \left[\left(H_{S_{f_i}^j}(k) - H_{S_{f_{i+1}}^j}(k) \right)^2 \Big/ max\left(H_{S_{f_i}^j}(k),\, H_{S_{f_{i+1}}^j}(k) \right) \right] \tag{4}$$

where$H_{S_{f_i}^j}(k)$is the bin value of the sub-image $S_{f_i}^j$ histogram. N is the overall number of sub-images, in our case is 16. K is the total number of direction bins; here it is equal to 5.

3. 3. Luminosity Degree (LD)

LD enables the distinction between the enlightened area and the dark ones by quantifying the brightness from the video frames. In general, the luminance degree can be computed by two different methods and using several models. In this work, the luminance of the different zones is compared by using the arithmetic average luminance; it is computed by finding the arithmetic mean of the luminance values of pixels in each zone. The pixel luminance value in a colour image is calculated by theYIK weighted sum models:

$$Luminance_{YIK} = 0{,}299\,R + 0{,}587\,G + 0{,}114\,B \tag{5}$$

where R, G and B are respectively the Red, Green and blue chromaticity value in RGB space. In a grey scale image, the luminance value is simply the pixel value.In contrast with the arithmetic average luminance method; the log-average luminance uses the geometric mean. Due the nature of the arithmetic operations, the choice of the arithmetic method is merely argued by the computation reduction and intuitiveness gain the relative luminance degree is extracted from each zone separately, in order to associate the values low, medium or high. Luminosity is thus calculated as follows:

$$L_i = \left(L.S - \sum_{j\neq i}^{Z} L_j.S_j\right)/S_i \tag{6}$$

where L and S are respectively the average luminance and the size (in pixels) of the entire image. $L_i$is the average brightness of zone *i* and S_i its size.

3. 4. People Existing (PE)

PE feature gives information about the presence of humans in the different zones. Both audio and visual modalities are exploited to get this information. At audio level, the presence of a non-modulated human speech is detected by analysing the spectra of sound recording. The presence of the fundamental and harmonic frequencies is explored taking into consideration that the range of the fundamental frequencies for the typical voiced speech is between 85 and 255 Hz (for both male and female).

3. 5. People Density (PD)

This feature approximately estimates the number of persons -if they exist – in each zone. The purpose of this part is to be able to detect, from the sound recording, the presence of one, two or more distinct speakers. The principle of the differentiation process is based on Independent Component Analysis (ICA). ICA is a statistical technique that aims to find a linear representation of a multidimensional random vector. This linear representation is made in the form of linear combination of non-Gaussian random variables (independent components) so that are maximally independent from each other (Hyvärinen2000).Consider, for example, a number of sound recordings $r_i(t)$ of *K* speech signals recorded by *M* distributed microphones. These *K* speech signals $s_k(t)$ are provided by *K* speakers simultaneously. The linear equations can represent each of the recorded signals as a weighted sum of the *K* speech signals.

$$r_i(t) = \sum_{j=1}^{K} a_{ij} \cdot s_j \qquad for\ all\ i = 1,..M \tag{7}$$

where $a_{ij}s$ are some parameters that depend on the distances of the microphones from the speakers. Assuming that the speech signals are statistically independent, the ICA allows the separation of the *K* original source signals from their mixtures $r_i(t)$ by estimating a_{ij}. FastICA (Hyvärinen 1999) is an efficient algorithm that is used in this system to perform the low computational estimation

4.Fusion And Recognition

The quantified semantic features that are extracted according to the detected depth zones must be processed to propose a classification. To this end, the Transferable Belief Model collects the extracted features from the different depth zones with the predefined information of each depth zone.

The transferable belief model (TBM) provides a model for the representation of quantified beliefs and for the combination of sources of evidence (Smets and Kennes1994). The TBM aims to determine the

real state C_k of the system using observations from five semantic features. The system state is supposed to take a number of exclusive context type hypotheses from the frame of discernment $\Omega_C = \{Ci, Co, Ce\}$ representing respectively the Cinema, Conference and Ceremony type. A mass function $m^{\Omega c}(A)$: $2^{\Omega_c} \rightarrow [0,1]$ with $\sum_{A \subseteq \Omega_C} m(A) = 1$ is defined on the power set of Ω_C.

The five semantic feature values from the three zones constitute the basic information and are fixed for each pair {Context type, Zone}. An expert knowledge basedmap is identified according to the normal properties of each context (Table.1). The interest of comparing the measured values with the reference ones is to extract the *matching degree* (Total T, partial P and Null N) according to number of coincidences. Thus the initial mass functions are assigned for each element of these frames of discernment: $\Omega_{match\_Ci} = \{T_{Ci}, P_{Ci}, N_{Ci}\}$, $\Omega_{match\_Co} = \{T_{Co}, P_{Co}, N_{Co}\}$ and $\Omega_{match\_Ce} = \{T_{Ce}, P_{Ce}, N_{Ce}\}$. With the conjunctive rule of combination (CRC), the initial mass functions derived from the five distinct sources (the semantic features) are combined. In order to collect information from these different frames of discernment, a common frame of discernment is built by applying the Cartesian product of the three sets. In this new space, the elements are triplets and the belief masses are re-allocated using the vacuum extension operation.Coarsening is a basic mechanism in TBM for space change. In particular, coarsening is exploited to transform the helpful information implied in the triplets into mass functions defined on the principal frame of discernment Ω_C.The final decision is made by choosing element with the highest pignistic probability.

Table 1. Audio Source Location (ASL) map. By the same way the four other feature maps are defined.

	Z1	Z2	Z3
Ci	High	Low	Medium ∪ High
Co	Medium ∪ High	Low	Low
Ce	High	Medium ∪ High	Medium ∪ High

5. Experimental Results

Due to the multitude of system parts and their different algorithms, the system stages are tested separately. In particular, the extraction method of each semantic feature is evaluated first, and then the entire context classification system is cross-validated. The experimental studies are conducted on a personal computer with a Core2 Duo 2.40 GHz processor, 4 GB RAM memory and MATLAB R2013a environment. For the visual semantic features (SRC and LD) and the whole DBCoM validation, the experiments are conducted on 30 real world video set with 400×200 size. The uncompressed videos are acquired by a video surveillance camera. The camera is fixed at the back of the hall. The dataset contains scenes for the same multipurpose hall while it is used for different targets; as a cinema room, as a conference hall and as ceremony hall. To evaluate the Screen Change Rate extraction part, a ground truth matching between existed screen changes and the automatic detected changes is manually constructed. We use the well-known performance parameters used in the most scene change detection methods; recall and precision (Table 4).

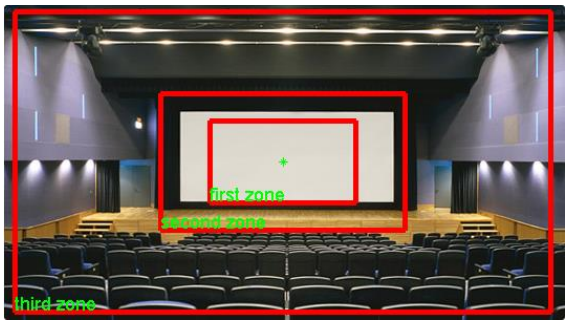

Fig. 3. A multipurpose hall scene that is segmented on depth zone by DERSS.

Table 2. Averaged Scene/Slide change detection evaluation.

	Average Rate of Change (changes/second)	Recall	Precision
Conference videos	0.12	0.79	0.89
Movie show videos	0.655	0.78	0.83

For the 2D localization of the source audio,results show the robustness of the method.The observed error is around 5cm between the theoretical position and the calculated position. This method error is not only very acceptable, but "absorbed" by the fact that it is unrealistic and too frustrated to consider a speaker as a single point. A second acceptable type of error was appeared; the intrinsic calculation error that is provided by Matlab operations. This negligible error is in the range of 9 mm.

Table 3. Average Error and RMSD for different tests of localization, for several sound types.

Sound Type	Average Error (cm)	RMSD
Applause	4.6590	0.0373
Speech	5.1537	0.0409
Choral	6.5426	0.0511
Music	6.3767	0.0498

For PE and PD, a pass rate of 77% is recorded. Failing tests appear where two people speak simultaneously and when the sound records contain interlaced voice and music.

The TBM prototype combines between Matlab and Excel to exchange tabulated data. The prototype supports the following conceptual tasks: manipulation of the different frames of discernment, managing the independency of certain parameters, creation of new frames of discernment and the product space, adaptation of CRC formulas and finally grouping different triplets following particular configuration.

Cross-validation (Refaeil zadeh 2009) is the statistical method that is used for evaluating the DBCoM system by dividing the data into two parts: one is used to learn or test model and the other is used to validate the model. Its most basic form is the k-fold cross validation. k = 10 is a good compromise. We examined 30 samples: a) 10 simulations of conference, b) 10 simulations of ceremony and c) 10 for simulation of cinema.Table 4 gathers the obtained results for each context type. Context classification presents very good performances in detecting the Cinema type with a sensitivity of 100%, a precision, a specifity and an accuracy > 80%, which demonstrates the robustness of the method in detecting this type. Table II indicates that a correct detection of two conference videos is missed. The first missed video is taken in poor lightning condition in comparison with the other conference scenes, so it is detected as a cinema. The second missed video is about non well-organized conference, therefore is biased to the ceremony decision. With regard to the ceremony type, it is simple to deduce from the low sensitivity rate (20%) the low capacity of the system on detecting this context type. However, the precision (70%) and specifity (95%) values show the fact that the poor performance on ceremony type detection does not affect the system behaviour in detecting the two other types (Cinema and Conference).

Table 4. DBCoM Performance indicators

Context type:	Cinema	Conference	Ceremony
Sensitivity	1.0	0.8	0.2
Precision	0.83	0.62	0.7
Specifity	0.8	0.75	0.95
Accuracy	0.87	0.77	0.67

6. Conclusion

This paper introduced a novel system that uses the spatial context modeling for context type classification. DBCoM (Depth Based Context Modeling) structures the derived knowledge from video-surveillance equipment according to the distribution of hall zones along the depth. After proposing an adaptive definition for "Context", DBCoM was built on the basis of a study that analyses the multimodal properties of each candidate context type in a multipurpose hall. An evaluation process for all system parts was applied separately on each module. Following the k-fold cross-validation strategy with differentperformance indicators we show that our system is robust overall, despite the poor results on detecting the 'Ceremony' context, due to some aberrant interpretations of PE and PD. Work is under progress to integrate visual information to PE and PD features. A possible solution would be through image processing, specifically with a shape detection to determine the number of present person in each zone.

References

Bremond F., Thonnat M. (1998). Issues of representing context illustrated by video-surveillance applications. In International Journal of Human-Computer Studies - Special issue: using context in applications, Volume 48 Issue 3, March 1998.

Charara N., Sokhn M., Jarkass I., Abou Khaled O. and Mugellini E. (2013). Dynamic Extended Rectangle Based Method for 3D Visual SceneSegmentation. International Review on Computers and Software (I.R.E.CO.S.), Vol. 8, N.4, April 2013.

Divvala S., Hoiem D., Hays J., Efros A., Hebert M. (2009). An empirical study of context in object detection. In Computer vision and pattern recognition, CVPR 2009. IEEE conference on, pp 1271–1278

Dourish P. (2003). What We Talk About When We Talk About Context. Journal of Personal & Ubiquitous Computing.

Hyvärinen A. (1999). Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Trans. on Neural Networks, 10(3):626-634.

Hyvärinen A., Oja E. (2000). Independent Component Analysis: Algorithms and Applications. Neural Networks, 13(4-5):411-430.

Marques O., Barenholtz E., Charvillat V. (2011). Context modelling in computer vision: techniques, implications and applications.Multimedia Tools and Applications, Springer.

Oliva A., Torralba A. (2001), Modeling the shape of the scene: a holistic representation of the spatial envelope. In International Journal of Computer Vision, Vol. 42(3):145–175.

RefaeilzadehP. , TangL., LiuH. (2009). Cross Validation. in Encyclopedia of Database Systems (EDBS), Editors: Ling Liu and M. Tamer Özsu. Springer, pp6.

Smets P. and Kennes R. (1994). The transferable belief model. Artificial Intelligence, 66:191-234.

Wolf, L. and S. Bileschi (2006). A Critical View of Context. International Journal of Computer Vision,69(2):251–261.