Context Identification from Video-surveillance Scenes of Multipurpose Halls

N. Charara[1,2], I. Jarkass[3], M. Sokhn[4], O. Abou Khaled[2] and E. Mugellini[2]

[1]EDST, Lebanese University, Beirut, Lebanon
[2]University of Applied Sciences of Western Switzerland, CH-1705 Fribourg, Switzerland
[3]LSI, University Institute of Technology, Lebanese University, Saida, Lebanon
[4]University of Applied Sciences of Western Switzerland, Techno-Pôle 3, CH-3960 Sierre, Switzerland
nour.charara@edu.hefr.ch, ijarkass@ul.edu.lb , maria.sokn@hevs.ch, Omar.AbouKhaled@ hefr.ch, elena.mugellini@hefr.ch

Abstract: **A novel method for multimodal context identification from video-surveillance footage of multipurpose halls is presented is this paper. After presenting a dedicated definition of 'Context' in computer vision systems, the goal is resumed by detecting the active context type among predefined ones. To this end, a spatial modeling of context is performed by extracting five discriminative semantic features according to depth zones. These zones are detected by depth-based scene segmentation method. These features are processed with the Transferable Belief Model (TBM) to propose a classification. Results show the validity of the method for context recognition.**

***Index Terms*—Context modeling, Pattern recognition, Transferable Belief Model, Video-surveillance.**

I. Introduction

A MULTIPURPOSE HALL 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. Many authors have used Context either implicitly or explicitly in their image understanding systems, but few have made the representation of context as central task, as we have proposed. The goal behind context modeling is to define formalism on the context description, in order to enable the identification of the current context in a given video.

II. System Overview

1) Concept and Architecture

The system architecture is depicted in Figure 1. Basing on this proposed definition: *"Context is the environment where an event can occur. The isolation of this event, from its environment, leads to another interpretation of this event"*, our context modeling and classification system is distributed on two phases. 1) Preparation phase including the depth-based segmentation of the multipurpose-hall scene and the extraction of five semantic features. 2) Classification phase where the mass functions are modeled by Transferable Belief Model operations.

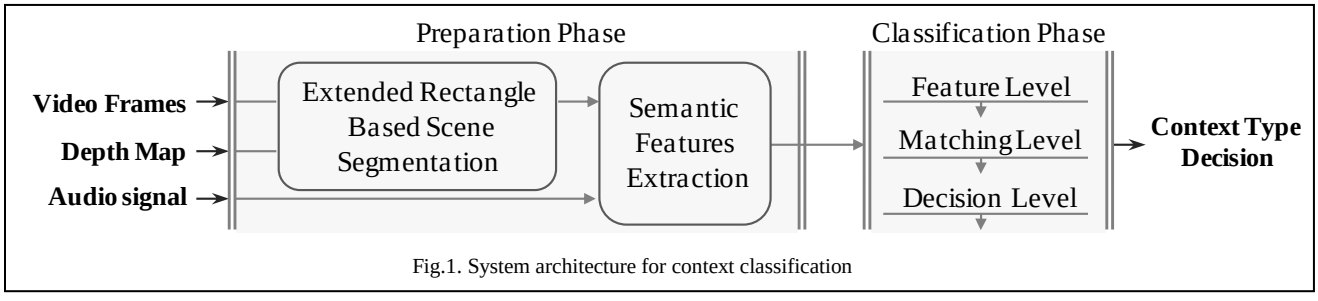

Fig.1. System architecture for context classification

The system takes three input types. In addition to the video frames and audio signals, the depth information is explored. The scene depth map is performed through a passive depth computation method

using a stereo vision system. Using DERSS (Dynamic Extended Rectangle based Scene Segmentation) [2], the distribution of the three different zones (the spectator seats zone, the sidewalk zone and the onstage with the wide screen zone) in a multipurpose hall is automatically captured without intervention from a human operator or a priori information.

2) Semantic Features Extraction

Conceptually, the following high-level information is used to differentiate between contexts: (1) **Screen Change Rate SCR** to obtain the rate of change of the filmed screen from the cropped video stream, by using edge detection based method [1]. (2) **Audio Source location ASL** identifies the zone(s) that is/are the origin of the audio source. Time Difference of Arrival (TDoA) method is applied on a three microphones array to compute the coordinates of the audio source. (3) **Luminosity degree LD** enables the distinction between the enlightened area and the dark ones by quantifying the relative brightness of each zone from the video frames. (4) **People Existing PE** gives information about the presence of humans in the different zones. Both audio and visual modalities are exploited to get this information. (5) **People Density PD** estimates the number of persons -if they exist – in each zone. ICA (Independent Component Analysis) based differentiation method is used to detect, from the sound recording, the presence of distinct speakers.

3) Fusion and Recognition

The Transferable Belief Model (TBM) collects the extracted sematic features from the different depth zones with the predefined information of each depth zone. The TBM aims to determine the real state C_k that is supposed to take one of the 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 is defined on the power set of Ω_C . This mass function that is assigned to each subset A of Ω_C undergoes to different TBM operations.

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 based map is identified according to the normal properties of each context. The *matching degree* (Total, partial and Null) is extracted 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 $\Omega_{match\_Ce} = \{T_{Ce}, P_{Ce}, N_{Ce}\}$.

The conjunctive rule of combination (CRC), the vacuum extension operation and Coarsening are basic mechanisms in TBM [3]. These operations are applied on the initial mass functions to collect information from the three frames of discernment of *matching degree.* The final decision is made by choosing element with the highest pignistic probability.

III. Experimental Results & Conclusion

An evaluation process for all system parts was applied separately on each module. The principal database consisted in 30 recorded videos that are taken from the same multipurpose hall where the purpose was to recognize cinema, conference or ceremony context type. Following the k-fold cross-validation strategy with different performance 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.

References

[1] N. Charara, I. Jarkass, M. Sokhn, O. Abou Khaled and E. Mugellini. "Tracking a Screen and Detecting its Rate of Change in 3-D Video Scenes of Multipurpose Halls," in *ICSIA 2013*, Barcelona, August. 2012.

[2] N. Charara, M. Sokhn, I. Jarkass, O. Abou Khaled, E. Mugellin. "Dynamic Extended Rectangle Based Method for 3D Visual Scene Segmentation," *International Review on Computers and Software (I.R.E.CO.S.), Vol. 8, N.4, April 2013*.

[3] Ph. Smets and R. Kennes, "The transferable belief model". Artificial Intelligence, 1994, 66:191-234.