

LAMV: Learning to Predict Where Spectators Look in Live Music Performances

Arturo Fuentes^{1,2}^a, F. Javier Sánchez¹^b, Thomas Voncina² and Jorge Bernal¹^c

¹*Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), 08193, Barcelona, Spain*

²*Lang Iberia, Carrer Can Pobla, 3, 08202, Sabadell, Spain*

Keywords: Object Detection, Saliency Map, Broadcast Automation, Spatio-temporal Texture Analysis.

Abstract: The advent of artificial intelligence has supposed an evolution on how different daily work tasks are performed. The analysis of cultural content has seen a huge boost by the development of computer-assisted methods that allows easy and transparent data access. In our case, we deal with the automation of the production of live shows, like music concerts, aiming to develop a system that can indicate the producer which camera to show based on what each of them is showing. In this context, we consider that is essential to understand where spectators look and what they are interested in so the computational method can learn from this information. The work that we present here shows the results of a first preliminary study in which we compare areas of interest defined by human beings and those indicated by an automatic system. Our system is based on the extraction of motion textures from dynamic Spatio-Temporal Volumes (STV) and then analyzing the patterns by means of texture analysis techniques. We validate our approach over several video sequences that have been labeled by 16 different experts. Our method is able to match those relevant areas identified by the experts, achieving recall scores higher than 80% when a distance of 80 pixels between method and ground truth is considered. Current performance shows promise when detecting abnormal peaks and movement trends.

1 INTRODUCTION


The commercial segment of the production of live performances is undergoing a transformation due to the growing automation of work tools. More frequently, television studios are incorporating the use of robotic cameras and replacing the camera operators for automation managers. This offers greater versatility and provides more complete motion control. In addition, its use also implies a saving in the costs related to the production, as a single operator can control several cameras.


This is different to what is being done currently, where a camera operator is required for each of the cameras installed. Nevertheless, these advances have a counterpart as the absence of human operators could lead to a loss of additional and subjective information that operators could bring to the filmmaker.


The scope of our work is the automatic live production of cultural events, in this case we focus on

music concerts. This kind of projects has already been tackled by other researchers in different domains, such as the works of (Chen et al., 2013) where they aim at the live production of soccer matches, similar to the system that Mediapro uses when broadcasting Spanish National Soccer League. Outside cultural events, the work of (Pang et al., 2010) deals with the automatic selection of the camera to show in video-conferences so the speaker is always on screen.

The domain in which our work can be enclosed has the advantage of having additional sources of information that can help in building the automatic system; we can integrate both video and audio analysis to determine which areas of the image should be shown in the main screen. It has to be noted that we aim to go beyond simple frame selection, our objective is to develop a system that can also generate a dynamic in terms of take selection, composition and length. Such a system, apart from its potential commercial future, allows us to generate knowledge related on how audiovisual narrative should work from a computational perspective which could be potentially extended to another domains.

^a  <https://orcid.org/0000-0002-1813-4766>

^b  <https://orcid.org/0000-0002-9364-3122>

^c  <https://orcid.org/0000-0001-8493-9514>

Related to this, we cannot forget the spectator, which is the one that receives the information; this information should contemplate both enuntiative and explicative components so the viewer can understand what is happening and the way the information is explained can generate in the spectator some kind of engagement. Our first approach, named Live Automation of Musical Videos (LAMV), deals with the enuntiative analysis of the action.

We rely on the use of texture descriptors along several patches of the image along several frames to search for those signatures that allows us to identify those image areas where big changes (associate to movements) are produced.

The objectives of this paper are two-fold: 1/ Identify where spectators look and 2/ Signature search. The first one deals with the acquisition of several annotations by experts and their integration to define which areas of the image are more relevant from an expert point of view. The second part is related with the analysis of texture information to define those signatures that can be attained to relevant image areas.

2 RELATED WORK

The first part of our system must deal with finding what is important in the image. In this case, we should not only look for what is relevant under a computational point of view; we also have to consider what would be important for a given person.

This is covered in the literature under the research domain known as visual saliency (Cazzato et al., 2020). The salience (also called saliency) of an item -be it a an object, a persona, a pixel, etc. -is the state or quality by which it stands out from its neighbors.

Saliency detection is considered to be a key attentional mechanism that facilitates learning and survival by enabling organisms to focus their limited perceptual and cognitive resources on the most pertinent subset of the available sensory data. When attention deployment is driven by stimuli, it is considered to be bottom-up, memory-free, and reactive. Conversely, attention can also be guided by top-down, memory-dependent, or anticipatory mechanisms. There are several computational saliency models already proposed in the literature, such as the works of Itti et al. (Borji et al., 2012; Itti et al., 1998), Tsotsos et al (Bruce and Tsotsos, 2009) or Seo et al. (Seo and Milanfar, 2009), just to mention a few of them.

In our case we must deal with the difficulty of defining which are the salient regions in an image where all elements (i.e. music performers) look alike. To solve this, we propose to incorporate the defini-

tion of visual differences between frames to our definition of saliency. The use of this type of saliency has already been proposed in the literature, as shown in the works of Rudoy et al. (Rudoy et al., 2013). In this case, the authors use the fixations from previous frames to predict the gaze in the new frame. The novelty here is the combination of within frame saliency with dynamics of human gaze transitions.

As mentioned before, we will also have to deal with temporal coherence of system output. The incorporation of temporal information has attracted the attention of several researchers of different domains, from the analysis of colonoscopy sequences (Angermann and Bernal, 2017) to tracking moving objects (such as driving lanes) in the development of assisted driving systems (Sotelo et al., 2004).

Regarding the domain of application, one of things to take into account is that the events we are dealing with are live. This means that, despite the fact that there is a lot of rehearsals (their amount depends on available resources) there is always a certain amount of improvisation associated with making decisions in very short periods of time and the necessity of adapting to unexpected events.

There are several available solutions in the sports domain such as the Multicamba system (Yus et al., 2015) used for instance in sailing to track the boats. With respect to soccer matches Mediapro has developed Automatic TV¹, a complete commercial solution that allows to select the best shot and send it directly to the viewers.

It has to be noted that the context we are dealing with presents some differences with the sports one. The technical director in front of a live concert has to integrate both video and audio information which adds more complexity to the task.

As LAMV will also integrate music channel processing, works regarding polyphonic music recognition (Gururani et al., 2018; Toghiani-Rizi and Windmark, 2017; Costa et al., 2017; Costa et al., 2011) score image recognition (Dorfer et al., 2016) natural synchronization of the visual and sound modalities (Zhao et al., 2018) should be taken into account.

3 METHODOLOGY

We expose in this section the two main parts of this contribution: the definition of the relevant areas of image as a result of the analysis of the income provided by several experts and the automatic extraction

¹automatic.tv

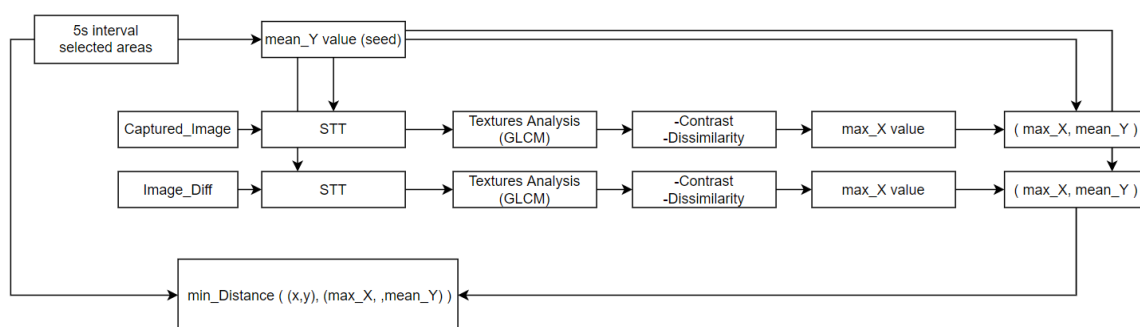


Figure 1: Flow chart of our first approach of LAMV, which shows the process followed to evaluate the minimum distance between the selected areas by the subjects with respect to our candidate.

of relevant features from the video sequence to estimate the key areas to look at.

The whole pipeline of our methodology can be observed in Fig. 1. As it can be seen, we have two different sources of data: selected areas by subjects and the input video. With respect to the former, we analyze a 5 seconds interval and we calculate the mean vertical value of the selections, which will be used as a seed in the image analysis module.

The second source of data is based on two modes of 5 seconds long input video -original frames and a difference of these frames (t, t-1) - whereby the two STT 125 pixels height are generated and analyzed through GLCM considering two descriptors: contrast and dissimilarity. The point candidate is represented by the maximum value of each descriptor -the X value -, and by the Y value - the pre-calculated mean value-. The minimum distance between the candidate and the closest ground truth candidate is considered to assess the performance of LAMV system.

3.1 Where Do Spectators Look at?

To gather as much useful information as possible, we ran an online experiment (available at <https://videoexp.dtabarcelona.com>) in which several experts were asked to observe different sequences (extracted from real music concerts) and indicate which image are was more relevant to them.

We established two different scenarios. The first one already proposed the participants five predetermined regions of interest whereas the second one allows them to freely browse the scene to search for the area that attracted their attention.

In both cases, the subject had to move a green box over the image and locate it over the area of interest, making the spectator act as a live producer. The volunteers were required to select a box (or a position) at least once every 10 seconds. Contrary to the automatic system, the participants also had the help of the audio channel to assist them in their decision. In

our case, we are only interested (as of now) in finding those visual cues that could also be attractive to the attention of the potential spectator.

The aggregation of the selections made by the individual experts is the basis to generate the ground truth which will be used to assess the performance of our method. In this case, we add each individual fixation as an individual box over the image and we select as the relevant area in the image the one in which more of these individual selections coincide.

3.2 Defining Signatures

Still frame analysis can be useful to detect which areas are more salient in an individual image but, in our case, we are more interested on those sudden changes in the image that could be associated, for instance, to an instrument that has just started to be played or any unprecedented happening in the plateau.

To build this automatic relevant region extraction module, we propose to use the so-called Spatio Temporal Texture (STT) slices. To generate them first we set a vertical position in the image, which can be used as reference. This position is set (in this experiment) as the mean y-value of the participant selections for this particular set of frames but the analysis that we present in this work could be easily extended for any given vertical position. Anyway, as selections fixations can vary over time (and they should do, as relevant events can appear in other image positions), we recalculate this vertical position each 5 seconds.

Once this vertical position is set, we generate next the slice by rendering the information of the pixels along the horizontal axis; the concatenation of this information over time generates the final slice. It has to be noted that we generate two types of images, depending on the source of information.

The first one directly renders the grey scale values of the pixels that share the same vertical coordinate; before doing this, contrast of the image is adjusted previously so the dynamic range of the image is re-

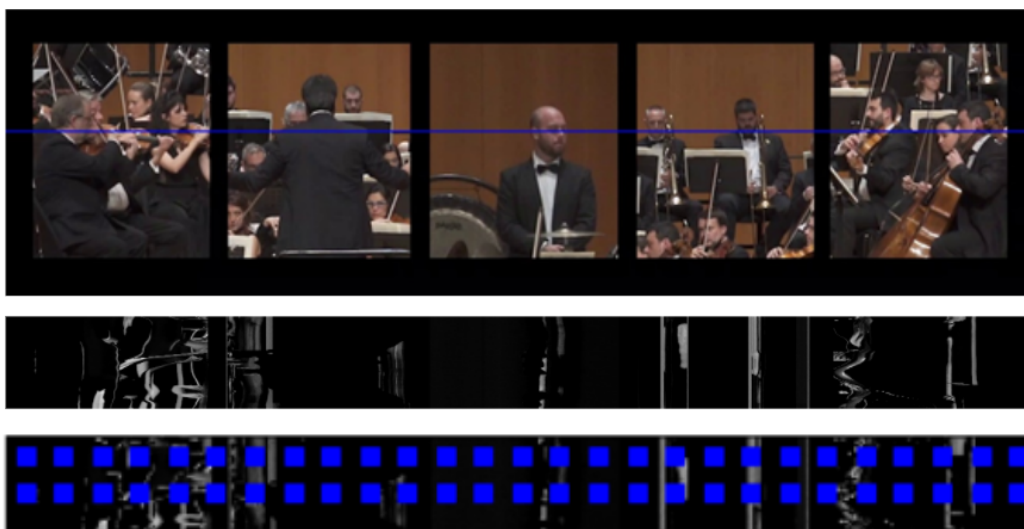


Figure 2: Graphical example on the automatic extraction of image patches. Image at the top shows the five predetermined regions that were shown to the volunteers. The blue horizontal line represents the mean value for the vertical coordinate of the selections over 5 seconds of video. The image in the middle represents the SST volume whereas the image in the bottom displays the 27 uniformly distributed patches over the volume.

duced. The second one uses as source information the difference between consecutive frames; apart from the previously mentioned contrast adjustment, Gabor filtering is applied to improve efficiency and motion fidelity (Hao et al., 2017). While our method has similarities to this work, we add as novelty the calculation of the slice image from difference of frames.

Considering these two different sources, we build our hypothesis over the fact that those image regions that present higher variation over time would correspond to the most interesting regions in the opinion of the volunteers. It has to be noted that, if we analyze the STT slices vertically, we are able to explore variations over time in a specific region in the image.

To assist us on the definition of image signatures we propose the use of Grey Level Co-occurrence Matrix (GLCM) (Gao and Hui, 2010), as it is meant to be useful on detecting information levels with high semantic content, which is the one we are looking for. In this case GLCM does not suffer from illumination changes as the lighting of the scene is uniform for the whole scene; we are dealing with pre-recorded musical concerts where the illumination conditions have been previously set and tested properly to allow a clear view of the music sheets by the musicians.

From the several descriptors that could be used, we have chosen dissimilarity and contrast as the ones that could lead us to easily determine areas with high change (and, therefore, to be prone to attract the attention of the viewer). It has to be noted that dissimilarity increases linearly with differences on grey level whereas contrast has an exponential increase.

We apply these descriptors over predetermined regions of the STT; in our experiments we have defined 27 different patches of size 50x50 that are spread equally over the slice. We calculate both descriptors for each of the patches and we select as final relevant region proposal the one where their maxima is achieved. We show in Fig. 2 a graphical example on how these patches are created.

4 EXPERIMENTAL SETUP

We validate our first iteration of the LAMV system by comparing the regions of interest that it automatically provides with the ones labelled by the participants in the online experiment.

Volunteers had two sets of sequences to analyze: one containing predetermined RoIs and another one where they could freely explore the image. The total length of the video was of 5 minutes and 30 seconds; spectators should indicate a region in the image at least once every 10 seconds.

It is important to mention that both sets were extracted from different music concerts, recorded at April 2019 and January 2020 at Auditori Sant Cugat and L'Auditori de Barcelona, respectively.

Each of these selections was saved with a time stamp. The ground truth generation experiment was performed with the support of p5js Javascript library. As of now, 16 different individuals have taken part in the annotation experiment. We show in Fig. 4 an example of ground truth creation for a given frame.

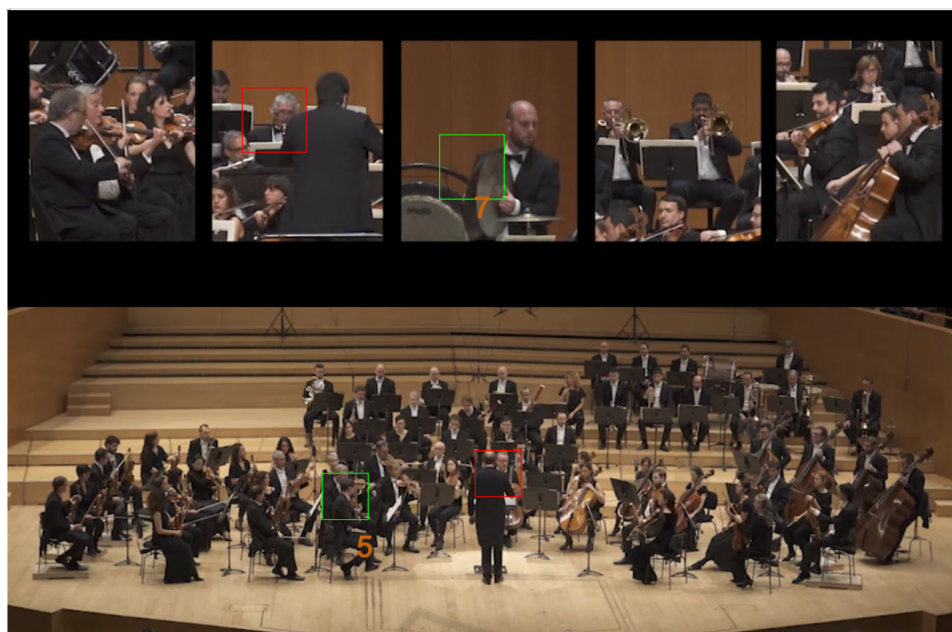


Figure 3: Example of the annotation experiment. Image at the top shows an annotation task where the user has to place the box (in green) over five predetermined boxes. The red box represents a selection by the user. The image at the bottom presents a full scene, using the same notation to display user selection.

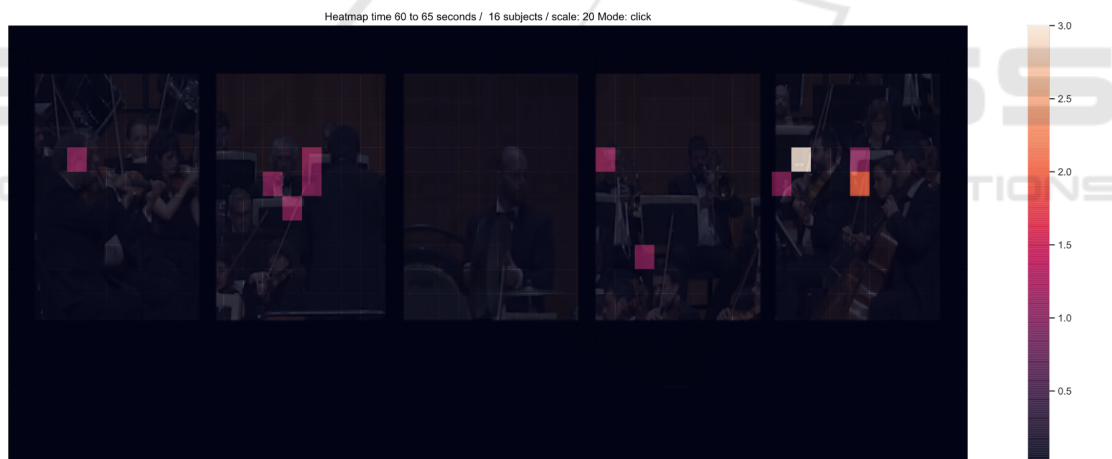


Figure 4: Example of ground truth creation from the annotation experiment.

We propose as validation metric the use of Recall, defined as the percentage of correct RoI detection over the total of RoIs in the sequence. We calculate the RoI provided by LAMV as follows: we select 125 lines of the STT slice which correspond to 5 seconds of the video sequence. Over this sub-volume, we define the before mentioned 27 50x50 patches and we calculate dissimilarity and contrast descriptors over them (using as the source image the original one and the difference of images).

The comparison of the values of the descriptors over these 27 patches provides a candidate. To assess

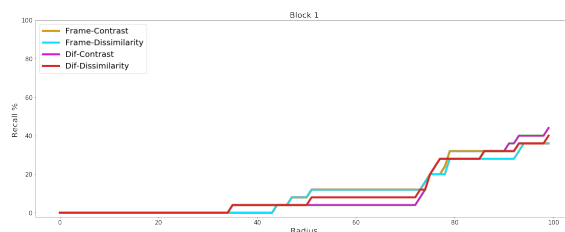


Figure 5: Performance of LAMV system over video sequences with predetermined Regions of Interest.

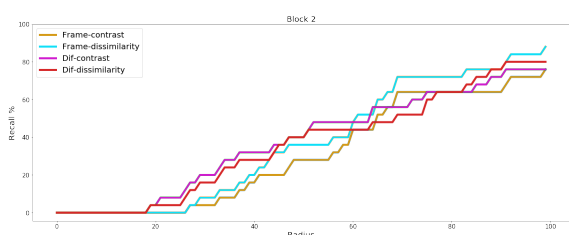


Figure 6: Performance of LAMV system over video sequences without predetermined Regions of Interest.

if this candidate correspond to the actual RoI defined by the expert, we calculate the euclidean distance between the centroids of the candidate and maxima of the ground truth for this particular set of frames. To label a detection as a True Positive we set a distance threshold that should not be surpassed.

Given that we analyze temporal frames of 5 seconds, the maximum number of TP over a video sequence will be equal to $max_T P = \frac{v_l}{5}$, where the video length (v_l) should be expressed in seconds.

5 RESULTS

Figures 5 and 6 shows overall performance of the system. Several conclusions can be extracted by the analysis of these graphs.

First, it appears that there is a big difference in performance associated to the way information is presented to the experts. Surprisingly the method is able to match better the ground truth when free selection of the region of interest is prescribed to the experts.

It has to be noted that the sequences were different in complexity, being easier to determine a sudden change in the scene when we have the whole picture. In the predetermined region analysis, a small change in a reduced region in the image weights more than in the full image, benefiting isolated sudden and small changes rather than actual changes in the scene.

This difference in complexity affects also the recall scores and the impact of the different descriptors. We associate this to the following: when dealing with predetermined regions where small changes affect more, the value of the contrast descriptor is more sensible to these slight changes as its value has an exponential profile whereas, for the analysis of the full images, these small changes tend to be mitigated.

We do not appreciate remarkable differences among descriptors and between image source though it is interesting to observe the increase of relevance of the dissimilarity descriptor when higher distances between centroids are allowed.

As can be seen from these results and, though the performance is promising, there are some cases in which our method fails to correctly determine the region of interest. We show in Fig. 7 an example of a patch which shows a trivial, routine movement of a wind instrument. In this case the system mislabels the preparation movement of the musician (just before starting to actually play the instrument) with an actual relevant movement in the scene. This shows that more effort should be made to incorporate semantics (and also to characterize particular movements such as the one shown in the example) in order to improve the performance of LAMV.

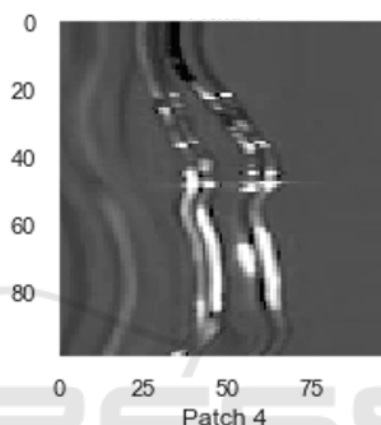


Figure 7: Example of one of the erroneous detection provided by the system.

We want to highlight some of the findings that we have observed by the analysis of the fixations provided by the volunteers.

Fig. 8 shows the difference in how selections are performed depending on the content of the video sequence, more precisely, depending on the intensity of the movements/actions that are happening in different time intervals during a same concert.

The blue distribution of selections is associated with an interval of the concert where the level of action is low whereas the orange distribution corresponds to a part of the concert when lots of movements are happening at once (for instance, a certain part of the musicians start to play the instrument).

The image on the left shows the amount of clicks during both periods; we can observe that the distributions present clear differences being the number of clicks higher for the more intense interval. By looking at the image on the right we can clearly make an association between the number of those clicks and the image area covered by them.

With respect to clicks occurrences, our hypothesis was that when the music is more intense, it is reflected with a more complex scene under a visual

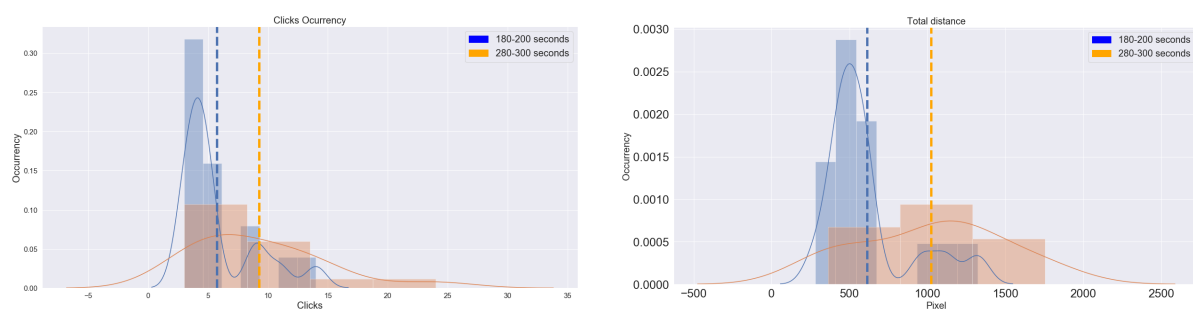


Figure 8: Example of the difference in the distribution of the number of selections and their coverage over the scene in two different time intervals.

point of view so a higher number of different areas are selected. We tested this using a paired sample t-test and the results confirmed our initial assumption, yielding a p-value < 0.05 .

Regarding selection distribution over the image, our hypothesis was that, apart from a higher number of selections, these selections would be more spread over the image. Again we tested this using a paired sample t-test and the results showed that, in fact, there are statistically significant differences between the two distributions (p-value < 0.05).

6 CONCLUSIONS

6.1 Main Conclusions

We have presented in this paper our first module of the LAMV system. This system aims to assist technicians when doing a live production of a music concert. The objective of the system is to indicate which areas of the image are more relevant (not only in terms of visual information) in order them to be broadcasted.

Though the complete system would incorporate audio information, key in this context, we wanted first to explore if the analysis of pure image information could be useful to indicate those relevant image areas.

To do so we have prepared a first experiment which compares the areas provided automatically by our system with those indicated by several volunteers that took part in an online experiment. LAMV system determines the area of interest by integrating image information over time using a STT volume and then extracting features from several patches.

Preliminary results show good correspondence between LAMV system and the ground truth though it also shows that semantic information should be added to improve its performance.

6.2 Future Work

With respect to the following steps in the development of the LAMV system, it has to be noted that the work in this article corresponds to one of the parts that compounds the LAMV system, specifically the pre-attentive analysis of the scene which is capable of detecting rapid changes that may arise on the scene as well as the unbalance between image areas, specially those with low activity.

At its current stage, the system works with using as starting seed for the analysis a vertical position calculated from the acquired ground truth. In the final system this seed will not be provided and several values for the vertical position will be considered.

Preliminary results show that LAMV is a good candidate to point out where to add more processing in specific areas of the scene but its current analysis is mainly based on changes over time. However, we cannot forget the role that audio information might have played in ground truth generation by experts. We have observed that experts tend to correctly select the areas where musicians are playing the predominant instruments, which can be clearly heard in the audio channel but that cannot be easily seen.

This clearly indicate us that part of the effort in the next steps should be devoted to the selection of descriptors with greater semantic feature structure which can help to locate the instruments even if they are not completely visible. This, for sure, would require of more advanced image processing techniques to perform a more accurate object detection.

In this context, we plan to study the incorporation of trending techniques such as those Convolutional Neural Networks integrating optical flow estimation such as FlowNet (Dosovitskiy et al., 2015). We would also like to explore the potential of recent methodologies such as LSTM or RNN, paying special attention to Transformers architecture and Vision Transformers for image recognition².

²<https://github.com/lucidrains/vit-pytorch>

Finally, we would also like to build a robust validation framework for the whole pipeline which would require acquiring and annotating more concerts.

ACKNOWLEDGEMENTS

This work was supported by the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya (SGR-2017-1669), by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) through the Doctorat Industrial programme and by CERCA Programme/Generalitat de Catalunya.

The authors want to thank Auditori de Sant Cugat and L'Auditori de Barcelona for granting us access to recordings of music concerts.

REFERENCES

- Angermann, Q. and Bernal, J. e. a. (2017). Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*, pages 29–41. Springer.
- Borji, A., Sihite, D. N., and Itti, L. (2012). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69.
- Bruce, N. D. and Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3):5–5.
- Cazzato, D., Leo, M., Distante, C., and Voos, H. (2020). When i look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking. *Sensors*, 20(13):3739.
- Chen, C., Wang, O., Heinzle, S., Carr, P., Smolic, A., and Gross, M. (2013). Computational sports broadcasting: Automated director assistance for live sports. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.
- Costa, Y. M., Oliveira, L. S., Koerich, A. L., and Gouyon, F. (2011). Music genre recognition using spectrograms. In *2011 18th International Conference on Systems, Signals and Image Processing*, pages 1–4. IEEE.
- Costa, Y. M., Oliveira, L. S., and Silla Jr, C. N. (2017). An evaluation of convolutional neural networks for music classification using spectrograms. *Applied soft computing*, 52:28–38.
- Dorfer, M., Arzt, A., and Widmer, G. (2016). Towards score following in sheet music images. *arXiv preprint arXiv:1612.05050*.
- Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. (2015). FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766.
- Gao, C.-C. and Hui, X.-W. (2010). Glem-based texture feature extraction. *Computer Systems & Applications*, 6(048).
- Gururani, S., Summers, C., and Lerch, A. (2018). Instrument activity detection in polyphonic music using deep neural networks. In *ISMIR*, pages 569–576.
- Hao, Y., Xu, Z., Wang, J., Liu, Y., and Fan, J. (2017). An effective video processing pipeline for crowd pattern analysis. In *2017 23rd International Conference on Automation and Computing (ICAC)*, pages 1–6. IEEE.
- Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.
- Pang, D., Madan, S., Kosaraju, S., and Singh, T. V. (2010). Automatic virtual camera view generation for lecture videos. In *Tech report*. Stanford University.
- Rudoy, D., Goldman, D. B., Shechtman, E., and Zelnik-Manor, L. (2013). Learning video saliency from human gaze using candidate selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1147–1154.
- Seo, H. J. and Milanfar, P. (2009). Nonparametric bottom-up saliency detection by self-resemblance. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 45–52. IEEE.
- Sotelo, M. A., Rodriguez, F. J., Magdalena, L., Bergasa, L. M., and Boquete, L. (2004). A color vision-based lane tracking system for autonomous driving on unmarked roads. *Autonomous Robots*, 16(1):95–116.
- Toghiani-Rizi, B. and Windmark, M. (2017). Musical instrument recognition using their distinctive characteristics in artificial neural networks. *arXiv preprint arXiv:1705.04971*.
- Yus, R., Mena, E., Ilarri, S., Illarramendi, A., and Bernad, J. (2015). Multicamba: a system for selecting camera views in live broadcasting of sport events using a dynamic 3d model. *Multimedia Tools and Applications*, 74(11):4059–4090.
- Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. (2018). The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586.