Scene Composition in Augmented Virtual Presenter System

Ting-Xi Liu[1], Yao Lu[2], Li-Jing Zhang[3], Zi-Jian Wang[4]

[1]*School of Computer Science, Beijing Institute of Technology, Beijing, China*
[2]*School of Computer Science, Beijing Institute of Technology, Beijing, China*
[3]*School of Computer Science, Beijing Institute of Technology, Beijing, China*
[4]*School of Computer Science, Beijing Institute of Technology,* [4]*China Central Television, Beijing, China*
[1]iutx@bit.edu.cn, [2]vis_yl@bit.edu.cn, [3]focus_zlj@163.com, [4]wangzijian@cctv.com

Abstract: In soccer match, presenters or commentators are needed to help the audiences understanding the match more clearly. For better visual effect, we design an Augmented Virtual Presenter System which can integrate presenter's image into the soccer field in match video. In fact, it is a scene composition process. In this paper, we will illustrate the structure and features of this system, and propose several solutions for the key problems. For scene composition, we design an algorithm consisting of automatic matting, localization, and occlusion processing. For occlusion problem, we propose a mixture solution including interactive and semantic segmentations for different scenarios.

1. Introduction

Soccer is one of the most popular games in the world. During the important soccer match events, such as FIFA World Cup or Euro Champions League, billions of fans will revel in it far into the night.

Most people enjoy soccer matches on TV because it is expensive and inconvenient for them to watch it in the stadium. The screen limits the information we could obtain from the match, so we need a commentator to narrate the game in the live. However, what a commentator could offer us is just voice. Obviously, it will be quite impressive for audiences if we let a presenter walk into the field, and describe the games or players more specifically and vividly.

For this purpose, we design an Augmented Virtual Presenter System. Virtual Presenter is a research hotspot in many fields. In [1] Noma et al create a virtual human presenter who accepts speech texts, and acts in real-time 3D animation synchronized with speech. [2] propose a method for modeling presentations in virtual environments, and [2] presents DynamicDuo, a system that uses an animated agent to help inexperienced speakers delivering their presentations in front of an audience.

Comparing with the systems mentioned above, it is not necessary for our system to model in 3D, and use real-human presenter instead. It will significantly reduce the complexity of the system, and enhance the reality of visual effect. Meanwhile, we propose a series of solutions to perform video synthesis automatically according to features of soccer match video.

The main functions of this system are matting presenter section in each frame of pre-recorded video, and inserting it into match video appropriately. We will illustrate the details of the system structure in next section.

There are two key problems in this system: scene composition and occlusion processing.

Scene composition: The match video and the presenter's video are obtained respectively. Therefore, how to synthesize scene with presenter's image and ensure the visual realism is the main problem to solve.

Occlusion processing: Generally, there are 22 players, 4 referees and other staffs in soccer field. So when the presenter is embedded in the ground, we need to deal with the occlusion problem in presenter's movements.

For scene composition, we propose a robust method to mat presenter from pure color background, then determine location and scale based on object detection using convolution neural networks.

For occlusion problem, we propose two kinds of solutions in different scenarios.

In close-up view, we use interactive image segmentation methods to segment the pixels of occlusion objects accurately;

In middle view or long view, we use semantic segmentation and matting methods to get coarse multi-object masks automatically.

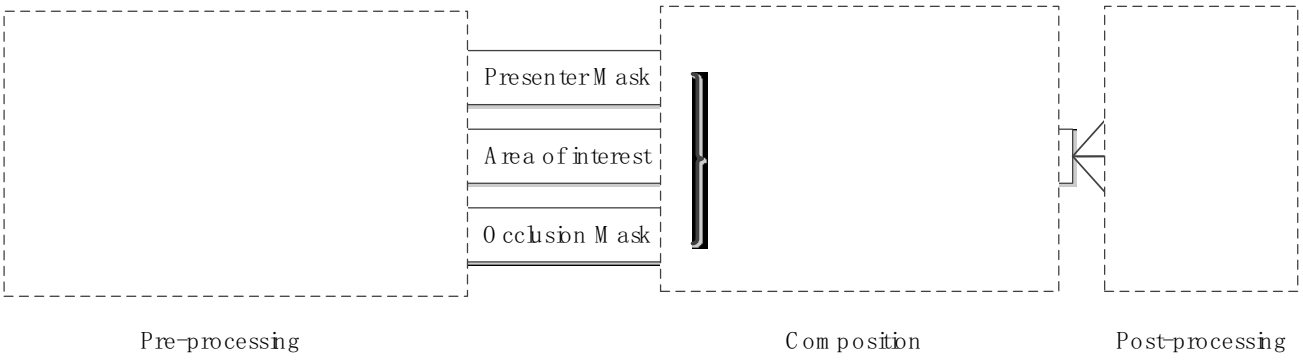

Figure1. Structure of Augmented Virtual Presenter System

In next section, we will introduce the structure of Augmented Virtual Presenter System. In section III and IV, we will illustrate the algorithms of scene composition and occlusion respectively. And section V will give a brief description of implementation details and results of our system.

2. Augmented Virtual Presenter System

The Augmented Virtual Presenter System (AVPS) integrates three modules, including pre-processing, composition and post-processing. The framework of AVPS is shown in Fig.1.

Pre-processing: This part accepts two video streams, presenter video and match video, as input.

For presenter video, we split it into frames and perform matting, store the masks and bounding boxes of presenter in each frame.

For match video, we perform localization and segmentation to obtain regions of interest and occlusion masks respectively.

Composition: With the information obtained above, we design a 3-layer composition hierarchy:

Layer-1: Specify scene from match video;

Layer-2: Insert the presenter's image into area of interest in Layer-1 using mask sequence from 1);

Layer-3: Cover the occlusion objects over Layer-2 by means of occlusion masks from 1).

Post-processing: In this part, we implement various of basic functions in video processing, e.g. video and audio editing (FFmpeg), simple animation effects (HTML5), and subtitle editing (FreeType).

3. Scene Composition

As we introduced in Section I, the match video and the presenter's video are obtained from different sources. The first key problem is how to merge the presenter's section into the scene in the match video appropriately.

For scene composition, we propose a two-step method:

Step1: Presenter image matting from pure color background to obtain a sequence of presenter's masks;

Step2: Localization based on object detection, to determine the coordinates and scales of presenter in match video.

3.1 Matting from Pure Color Background

There are many different matting approaches, e.g. Bayesian Matting [4], Poisson Matting [5], Closed-form Matting [6] and Robust Matting[7]. But these solutions cannot be applied in real-time video matting because of the complexity and demands of user interaction.

In order to obtain the presenter's section from video sequence in real-time automatically, we use a simple pure color background matting method.

We take several presenter's videos in front of a blue curtain background, which is not pure blue indeed. So we set a threshold, which is about 29~35 in our experiments. For each pixel in RGB color space, if both the value of blue component minus green and red are greater than the threshold, we consider it as a blue background point. After mask extraction by matting, we perform a Gaussian blur operation on the mask to smooth the border.

Algorithm 1 Pure Color Background Matting

Input: An image $\boldsymbol{I}$ with presenter in blue curtain background
parameter $\boldsymbol{Thres}$ is the threshold of background pixels
Output: Presenter's mask $\boldsymbol{Mask}$

```
function PureMatting(I, Thres)
    for each pixels I_i in I do
        [R, G, B] ← I_i
        if (B − R) > Thres and (B − G) > Thres then
            Mask[i] ← 255
        else
            Mask[i] ← 0
        end if
    end for
    Mask ← GaussianBlur(Mask)
    return Mask
end function
```

The algorithm and result of matting are given in algorithm1 and Fig. 2.

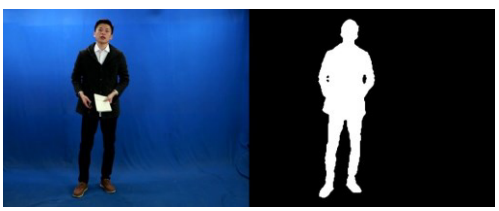

Figure 2. Result of Pure Color Background Matting

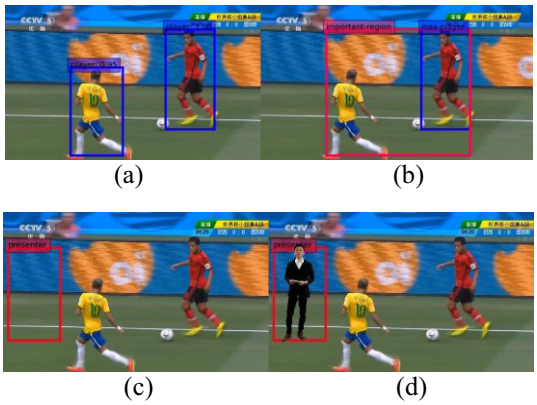

Figure 1. Procedure of localization. (a) shows the result of detection, (b) shows the region of interest (red bounding box), (c) shows the result of presenter's location and scale, (d) shows the composition result.

3.2 Localization By Object Detection.

In the first step we get a sequence of presenter's masks, and what's next is to insert the presenter into match video using these masks. Before that, a localization operation is needed to determine the locations and scales of presenter.

In order to perform localization automatically, we propose an algorithm based on object detection using convolutional neural network, to select the proper region for presenter according to the location of detected players. The localization operation consists of 3 steps:

Step1: Player detection.

Firstly, to get the players' attributes including locations and scales, we use a SSD (Single Shot MultiBox Detector)[8] model to perform detection. SSD is a method for detecting objects in images using a single deep neural network, it eliminates proposal generation procedure such as selective search in R-CNN[9] and region proposal network in Faster R-CNN[10]. This feature makes SSD easy to train and to be integrated into our system.

Step2: Selecting region of interest.

Based on the players' locations and confidences fetched in detection, we calculate the smallest bounding region that includes all important players, whose detection confidence is greater than a threshold (0.7 in our experiment).

Step3: Presenter localization.

In order to avoid crowding in visual effect, the presenter will enter from the side which is further to the region of interest. For example, in Fig. 3, the presenter enters the field from left. Then we set the corresponding edge (left in Fig. 3) as the far end of presenter's location, and determine the scale by the height of the player whose detected area is maximum in the image.

An example of the entire procedure is presented in Fig.3.

4. Occlusion Processing

As we know, in a soccer match, there are players, referees, staffs and other objects in the field. Our system attempts to insert a presenter into the ground, and the presenter need to enter from the outside of the shot.

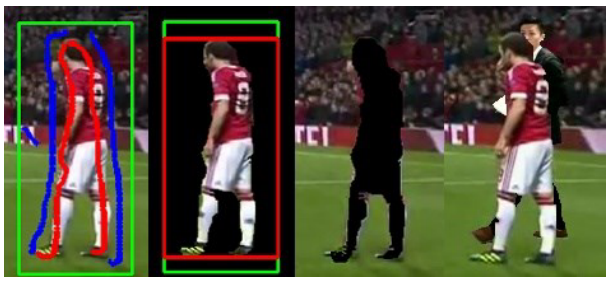

Figure 2. Accurate occlusion using interactive segmentation (GrabCut)

During the movements, the presenter should be occluded by some real players and objects in some case. Here we collectively define them as "occlusion objects".

For occlusion problem, we propose a framework which includes two processing method used in different scenarios:

Accurate extraction of occlusion objects for close-up view by interactive segmentation methods.

Automatic construction of occlusion hierarchies by pure matting or semantic segmentation for middle/long view.

4.1 Accurate Extraction of Occlusion Objects for Close-Up View

For close-up view, occlusion objects are too large to ignore the details in the border. In this case, we consider to apply interactive segmentation methods in accurate occlusion processing, e.g. GrabCut[11], Random Walk[12][13][14]. The result of accurate occlusion using GrabCut is shown in Fig. 4. This method acts better in the details compared with the automated approaches, but demands considerable user interactions.

4.2 Automatic Construction of Occlusion Hierarchies for Middle/Long View

For middle or long view, occlusion objects who occlude presenter are smaller. Under the case of low resolution, the details of these players and objects have little effect on visual effect. Meanwhile the number of objects are much more than that in close-up view, it's unreasonable to segment each objects manually by interaction methods.

The framework we propose for this case consists of two steps:

Step1: Generate complete, coarse masks of all occlusion objects in the field automatically;

Step2: Construct the occlusion hierarchies by footprint point.

Here we offer two alternative methods in step 1: 1) matting method, and 2) semantic segmentation using fully convolutional networks FCNs.

Method 1) is simple and efficient, can be used in scenarios that all the occlusion objects are in the green field area. It consists of 4 steps:

Step1: Use pure background matting to get green area as field;

(a) (b) (c)

Figure 3. Comparison of Matting method and Semantic Segmentation. (a) is the source image, includes 11 players, 1 referee and 1 staff. (b) shows the result of matting method. (c) shows the result of semantic segmentation using DeepLab model.

Step2: perform morphological transformation to remove the noise;

Step3: Find maximum contours, and fill the convex hull;

Step4: Perform pure matting in maximum contours to get occlusion masks of players.

The result of method 1) is shown in Fig.5(b).

Obviously, matting method is poor in robustness when the occlusion objects are not in the field entirely. For example, the goal keeper in Fig.5(a), whose body is not involved in green field, cannot be extract by matting method. So we consider applying semantic segmentation based on convolutional neural networks to extract masks of occlusion objects.

FCNs[15] is a method for semantic segmentation. It builds "fully convolutional" networks that take input of arbitrary size and produce correspondingly-sized output by transferring pre-trained classifier weights, fusing different layer representations, and learning end-to-end on whole images.

While what we used in this system is the improved approach DeepLab[16], which proposes atrous spatial pyramid pooling (ASPP) to robustly segment objects at multiple scales, and improves the localization of object boundaries by combining the responses at the final DCNN layer with a fully connected Conditional Random Field.

We integrate the DeepLab model in our system, considering all the players, referees and staffs as a semantic type 'people', and performing automated player segmentation. The result is shown in Fig.5(c).

Fig.5 shows the comparison of two methods. As we can infer from the result, the matting method acts slightly better than semantic segmentation in accuracy, but fails in matting objects who are not in the green field area (e.g. the goal keeper in Fig.5(a)), while semantic segmentation could roughly segment all the people in the image. So we offer both of these two alternatives, meanwhile manually fine-tuning operations are available too.

By means of occlusion mask obtained above, we can find contours to mark the location of each players and objects. Here we defined a footprint point for presenter and each objects, which equals the bottom point of their contours. If the footprint point of an object is lower than presenter's, we consider it as an occlusion object. The procedure is illustrated in Algorithm 2, and an example of automatic occlusion processing is shown in Fig.6.

5. Experiment

The main framework of the system is developed in C++, while modules using deep learning methods are based on Caffe and its Python APIs. Other development details are shown in Table 1.

We test our system in a Nvidia GTX Titan X GPU, with the support of GPU acceleration we could integrate deep learning models into the system. In detection task, the SSD300 (VGG16) model could reach 46 fps, and in semantic segmentation it takes about 1.54 seconds to perform segmentation in the resolution 1280×720.

Algorithm 2 Automatic Occlusion Processing

Input: Original image $\boldsymbol{Match}$, $\boldsymbol{Presenter}$
footprint point $\boldsymbol{Foot}$ of presenter
Output: Composed image $\boldsymbol{Result}$

function OCCLUSION($Match, Presenter$)
  $Mask_{all} \leftarrow$ SemanticSegment($Match$)
  $Contours \leftarrow$ findContours($Mask_{all}$)
  for each contour $\boldsymbol{cnt}$ **in** Contours **do**
    if $Foot_{cnt}.y > Foot_{pre}.y$ **then**
      $drawContours(Mask_{occ}, cnt)$
    end if
  end for
  $Layer1 \leftarrow Match$
  $Layer2 \leftarrow Presenter$
  $Layer3 \leftarrow Match(Mask_{occ})$
  $Result \leftarrow Merge(Layer1, Layer2, Layer3)$
  return $Result$
end function

Fig.6 shows a typical scenario of presenter with occlusion hierarchies. The two players whose footprint points are lower than presenter's are considered as occlusion objects.

Table 1. Implementation details of AVPS

Acceleration	CUDA 8.0
GUI	Qt 5.7
Image Processing	OpenCV 2.4.10
Audio	FFmpeg
Subtitle	FreeType 2.6
Animation	HTML5

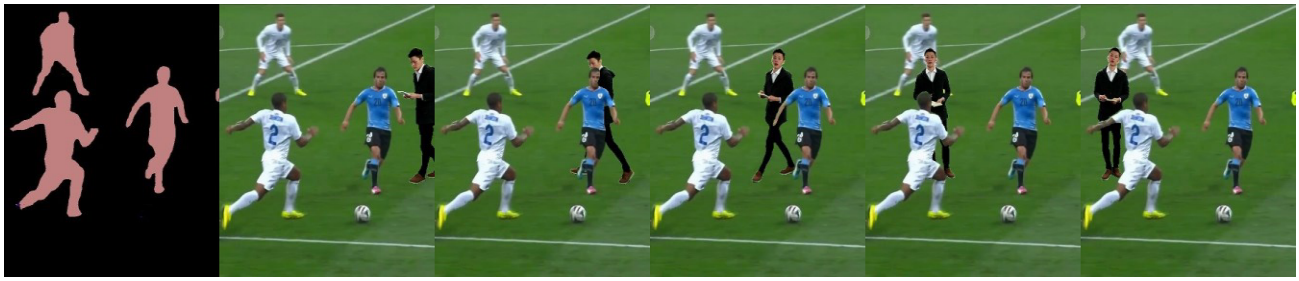

Figure 4. A typical scenario of presenter with occlusion hierarchies

6. Conclusion

We develop an Augmented Virtual Presenter System and implement most of the modules and features. During the development, we propose a series of methods to perform video composition, and to solve key problems like occlusion. We apply matting, object detection, interactive and semantic segmentations for better visual effect and user experience.

The main contribution of the system is that we make it highly automated to present soccer match video. In the whole system, excepting for basic editing operations, manual inputs that we need to offer are just 1) interactions in accurate occlusion objects segmentation, and 2) manual adjustment to fix defects or refine the locations, etc.

Besides, most of the components in our system could be performed in real-time, which makes the system applicable in either live stream or analysis after the match.

Our system is developed for soccer match, but with the support of deep learning methods in detection and segmentation, we could easily transfer the system for other sports video such as volleyball, basketball, American football and so on, and many types of TV programs which need a presenter.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61273273).

References

[1] Noma T, Zhao L, Badler N I. Design of a Virtual Human Presenter[J]. Computer Graphics & Applications IEEE, 2000, 20(4):79-85.

[2] Nijholt A, Welbergen V H, Zwiers J, et al. Introducing an Embodied Augmented Virtual Presenter Agent in a Virtual Meeting Room[J]. Acta Press, 2005.

[3] Trinh H, Ring L, Bickmore T. DynamicDuo:Co-presenting with Virtual Agents[J]. 2015:1739-1748.

[4] Chuang Y Y, Curless B, Salesin D H, et al. A Bayesian approach to digital matting[C]// Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, 2001:II-264-II-271 vol.2.

[5] Jian S, Jia J, Tang C K, et al. Poisson Matting[J]. Acm Transactions on Graphics, 2004.

[6] Levin A, Lischinski D, Weiss Y. A Closed-Form Solution to Natural Image Matting[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2008, 30(2):228-242.

[7] Wang J, Cohen M F. Optimized Color Sampling for Robust Matting[C]// Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. IEEE, 2007:1-8.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. 2015.

[9] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// Computer Vision and Pattern Recognition. IEEE, 2013:580-587.

[10] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015:1-1.

[11] Rother C, Kolmogorov V, Blake A. "GrabCut"[J]. Acm Transactions on Graphics, 2004, 23(3):309.

[12] Ju W, Xiang D, Zhang B, et al. Random Walk and Graph Cut for Co-Segmentation of Lung Tumor on PET-CT Images.[J]. IEEE Transactions on Image Processing, 2015, 25(3):1192-1192.

[13] Dong X, Shen J, Shao L, et al. Sub-Markov Random Walk for Image Segmentation[J]. IEEE Transactions on Image Processing, 2016, 25(2):516.

[14] Chefd'Hotel C, Sebbane A. Random Walk and Front Propagation on Watershed Adjacency Graphs for Multilabel Image Segmentation[C]// IEEE, International Conference on Computer Vision. IEEE, 2007:1-7.

[15] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4):640-651.

[16] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. 2016.