Improved Hockey Rink Localization via Augmentation and Temporal Frame Analysis

Jia Cheng (Jason) Shang — Vision and Image Processing Lab, University of Waterloo
Mehrnaz Fani — Vision and Image Processing Lab, University of Waterloo
David Clausi — Vision and Image Processing Lab, University of Waterloo
Mohammad Javad Shafiee — Vision and Image Processing Lab, University of Waterloo
Email: {jcshang, mfani, dclausi, mjshafiee}@uwaterloo.ca

Abstract

Deep-learning based hockey analysis generally requires automatic rink localization from broadcast videos. This information is used to determine the locations of players and the puck, which is important for further analysis such as puck trajectory and player behaviour. Models for this task determine the homography matrix used to warp the frame onto the rink template, or vice-versa. However, training models with good performance is challenging due to lack of training data. Augmentation algorithms have been shown to be effective for different machine learning tasks. Here we propose a set of new augmentation techniques specifically for the task of homography estimation to improve the model's reliability in new situations. To further improve smoothness and reliability of localization, we take advantage of refined homography between successive frames subsampled from videos in the inference stages. Results show that the new augmentation technique along with the smoothing approach can improve the performance by $\sim 2\%$.

1 Introduction

Hockey is an exciting fast-paced sport that is watched by millions of people, and each team constantly strives to improve and outplay their rivals. In order to study and improve player performance, teams, coaches, and analysts glean information from watching the players interact with other players and with the puck. With the advent of deep learning, it is now possible to automate parts of this analysis.

An important step of such analysis involves the identification of player and puck location. As hockey broadcast videos often do not have camera parameters, such location information must be identified automatically from the video feed itself. Thus, rink registration is performed to calculate how pixels in video frames map to the overhead view of the rink. An example can be seen in Fig. 1, which shows how an image looks when warped to the overhead view, and how the overhead template looks when warped onto the image.

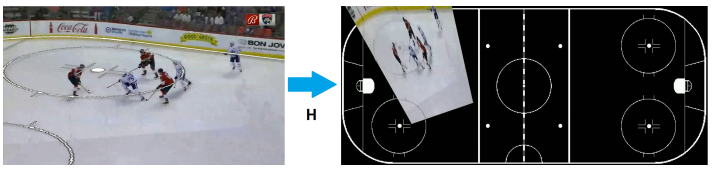

Fig. 1: Example of warping a video frame onto the overhead rink template (and vice versa) using homography.

Most models analyze each frame and output information used to calculate the homography warp between the overhead rink and the image frame. However, acquiring ground truth data for this task is difficult and expensive, and many previous papers do not have publicly released datasets.

In this paper, we hypothesize that stronger augmentation approaches can improve the model reliability and accuracy on limited training data. We also analyze the effect of using a new inference method to calculate the homography warp between successive frames subsampled from video. This adds temporal information, reduces shakiness, and improves reliability in difficult situations.

2 Related Works

Sports field registration is a vital component for analysis in many different sports, and thus different methods have been developed to perform this task effectively for sports such as soccer, basketball, tennis, and hockey. In deep learning, the models are trained on video frames from games in order to perform a homography estimation that can warp the frame onto the overhead field template, or vice versa.

Traditional homography estimation methods rely on matching features from pairs of images via methods such as SIFT [1] and ORB [2], before being used in methods such as RANSAC to calculate the homography [3]. DeTone et al. were one of the first to utilize deep learning to estimate homography, by using a VGG based model to estimate the location of 4 corners of one image in the image space of the other [4]. This can then be converted into a homography via techniques such as Direct Linear Transform (DLT) [5] [4].

Since then, many other models have been built for the purpose of homography estimation, and this includes various models specialized for sports field registration. Homayounfar et al. use semantic segmentation in order to isolate field marking information to use in a Markov Random Field [6]. Chen and Little set up a camera pose database with predefined poses, and then they select the best one compared to features and edge images extracted from the input image, before refining it [7]. Sha et al. also use a dataset method, except they use semantic segmentation output rather than edge images as the input involved in the comparison [8].

Nie et al. use a U-net based approach to estimate the location of a set of uniformly spread keypoints, which is then used to calculate the homography [9]. This keypoint based estimation is then refined based on feature heatmaps extracted from the image, alongside the previous frame's heatmaps [9]. Chu et al. build upon this approach by replacing the refinement step with better keypoint estimation based on dynamic filter learning [10].

Jiang et al. use a deep neural network to estimate the locations of 4 points of the input image onto the sports field template [11], in a similar manner as [4]. This produces an estimate which is used to warp the template and act as part of the input for a refinement network, which calculates the relative homography between the original image and the initial estimate [11]. Shi et al. improve this approach by warping their dataset images to generate synthetic images to improve their refinement model [12].

Our model is built upon on the one described in Shi et al. [12] with some changes and improved augmentation approaches to increase accuracy and reliability of the homography estimation.

3 Method

Our method for homography estimation is a two step approach. First we use an initial estimator to calculate an initial estimate via the four-point approach, which was first described in [4]. This initial homography is then used to warp the overhead template and is then fed into the refinement model alongside the original input frame. The refinement model is similar to that in [12], where two separate branches take in each input. The final output is the location of four points from one image in the image space of the other, allowing us to calculate the refinement homography needed to transform the initial estimate to be closer to the true homography. This refinement process can be iterated for improved results. Note that we omit the score branch and instead iterate a fixed number of times for simplicity. A diagram showing this process can be seen in Fig. 2

The initial estimator and refinement model were trained separately, using AdamW optimizer, smooth L1 loss, an initial learning rate of 0.0001, and weight decay of 0.3. There is a lack of publicly available datasets for hockey, so we use a 4501 image training set provided to us from Stathletes.

For training the refinement model, we generate synthetic data from our initial training in a similar manner as [12]. To generate a synthetic image, a rectangle of 4 points is taken on an original dataset image, and each point is perturbed. The resulting perturbed points and the original 4 points can be used to setup a homography matrix to warp the image to simulate zoom and translation.

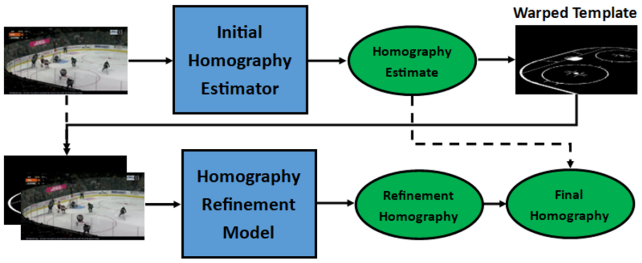

Fig. 2: Pipeline of the process, showing the initial estimator and the refinement model. The iteration of the refinement model has been omitted for clarity.

In our case, we make some changes during this process compared to Shi *et al.* [12]. We use the homography matrix generated from the perturbation along with the ground truth matrix of the training data to warp the template to use as the edge image training input, while keeping the original video frame as the frame input. We also alter the perturbation amounts due to our images being of a different size. This process can be seen in Fig. 3.

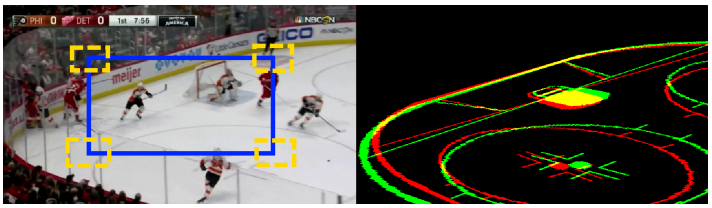

Fig. 3: Example Pertubation. On the left image, the blue rectangle shows an example set of 4 point chosen and the yellow rectangles show the possible perturbation locations. The right image shows how the original edge image (green) is perturbed into a synthetic edge image (red). The model is now fed the red perturbed edge image and the video frame, and tries to calculate the homography warp between them.

Additionally, rather than only generating 8 synthetic images per original dataset image as in [12], we dynamically generate a different synthetic edge image each epoch, to better simulate the vast potential rink orientations that are possible in hockey.

Our main contribution will be described in the following two sections, describing the augmentation improvements and the temporal video analysis we use.

4 Augmentation Improvements

Having a wide variety of training data is important to ensure that models don't overfit. However, preparing ground truth data is an expensive and time-consuming task. For this reason, various augmentations should be used to extend the training dataset, and thus improve the model's ability to generalize and improve results on new data.

As mentioned earlier, our process generates new synthetic edge images from each training set image each epoch. Before this process however, we also augment each training image via strong zoom, flip, and color augmentations. The homography warp needed to perform the zoom and flip augmentation is also calculated and used to alter the ground truth homography to ensure that it still matches the new augmented image. This is done randomly each time, and thus, coupled with the synthetic data generation, ensures that the model is trained on an extremely diverse training set. This added augmentation proved to be especially useful in cases where the camera zoomed in, which was previously a difficult case to manage. The accuracy increase these augmentations added can be seen in Table 1. Note that although the baseline is based on [12], dataset differences and potential differences in implementation make it difficult to perform a direct comparison.

We also utilize our own variant of copy-paste augmentation to further augment our images. The original copy-paste augmentation described in [13] enhances instance segmentation training by copying instances from other images and pasting them onto the current image, while altering the ground truth segmentation to account for the newly added instances. In our case, we copy instances of players from one image onto another to simulate natural occlusion of rink features such as lines, and to further augment the training data to improve model generalization. Current results show an increase in accuracy in both the initial estimator and the refinement model thanks to this augmentation, as seen in Table 2.

Table 1: Accuracy increases due to added homography and color augmentation on refinement model.

Model	IoU (part)	IoU (whole)
Refined Model (base)	96.9%	86.4%
Refined Model (with augmentation)	**97.1%**	**87.8%**

Table 2: Accuracy increases due to copy-paste augmentation on both initial estimator and refinement model.

Model	IoU (Part)	IoU (Whole)
Initial Estimator (no copy-paste)	96.0%	86.2%
Initial Estimator (with copy-paste)	**96.2%**	**86.3%**
Refined Model (no copy-paste)	97.2%	87.9%
Refined Model (with copy-paste)	**97.3%**	**88.1%**

5 Temporal Video Frame Analysis

Another aspect analyzed was the prospect of using the refinement model to calculate the homography between different frames in video. The refinement model is trained to identify the homography between a perturbed edge image and a video frame. The edge image from a previous frame can be considered to be similar to a perturbed edge image as long as the sampling rate of the frames from the video is reasonably high. Thus, we can use a different inference method where the refinement model is used to identify homography between successive subsampled frames from a video. This inference process can be seen in Fig. 4, where the refinement model can be used to calculate homography for each subsequent frame.

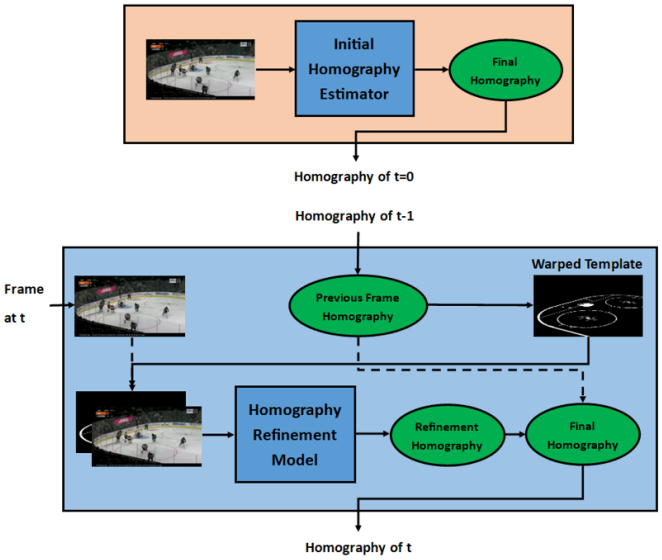

Fig. 4: Temporal video inference method. After the first frame homography is calculated, we can use the refinement method on each previous frame to calculate the homography needed to get from the previous frame to the current frame.

The benefits this approach offers include inference speed. Unlike the previous approach which required both the initial estimator and the refinement model for each frame, this method only requires the refinement model after the first frame. Furthermore, this introduces temporal information to the process to improve smooth-

ness and reduce shakiness of the resulting homography output for videos, as the model directly infers the difference between frames.

Finally, this improves reliability. The problem the initial estimator solves involves a much larger warp, as it involves warping from camera view to the overhead view. This is a harder problem to solve, and thus is more susceptible to unseen cases before. An example can be seen in Fig. 5, where the usual initial estimator plus refinement model is unable to estimate the homography well. This is due to the the initial estimator being unable to calculate a good homography for this case, and the refinement method being unable to fix the error as it is too far off. The refinement model in the temporal method solves an easier problem of just calculating the homography warp between the previous frame and the current one, and doesn't need to rely on the initial estimate for this case. Thus, in the temporal method we skip the initial estimator when dealing with this difficult frame, as we can compare the current frame with previous frames instead. Because of this, the temporal method is not as affected by this case.

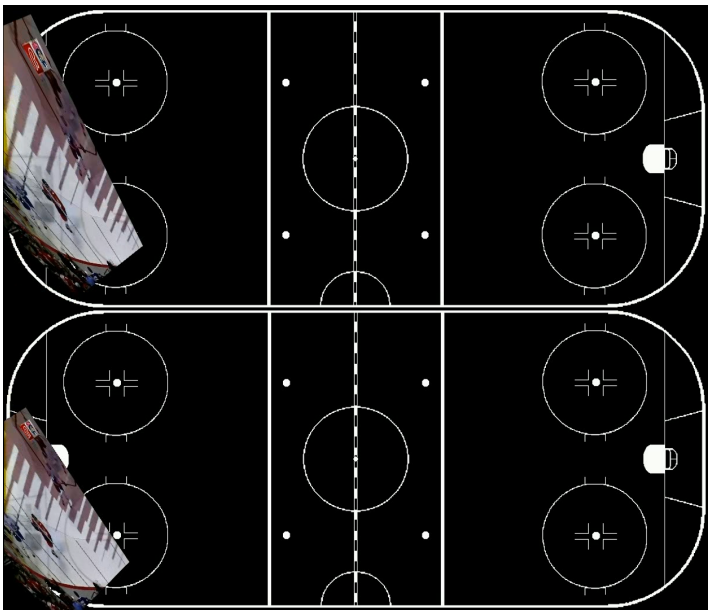

Fig. 5: Example where the temporal frame method is more reliable. The top shows a case where the initial estimator is unable to calculate the homography well due to the zoom and the visual effect covering part of the screen. The bottom image shows how the temporal frame method is unaffected as it calculates the warp from the previous frame to the current one, which is easier to do.

6 Future Work

Future work with this line of research would be to intensively investigate the effects of these augmentations and the synthetic data generation with a more thorough ablation study. There is likely further optimization that can be made regarding these techniques used to artificially increase the dataset for better model generalization.

As well, more development on the temporal frame model could be done to potentially take into account more that just the previous frame. For example, several past frames can be used, along with their homography information if available, in order to enhance the homography estimation for the current frame.

7 Conclusion

We have investigated the benefits of adding stronger augmentations in the training process of hockey rink localization models in order to improve generalization. We have also investigated the effects of running a refinement model to calculate the homography between successive frames subsampled from video, in order to produce smoother results and handle more difficult cases. The results have increased accuracy, and the qualitative effects can be seen in the model's ability to better handle zoomed-in frames and other more difficult to handle cases. Further work can be done to optimize this approach even more.

Acknowledgments

This work was supported by Stathletes through the Mitacs Accelerate program and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[3] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[4] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," *arXiv preprint arXiv:1606.03798*, 2016.

[5] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[6] N. Homayounfar, S. Fidler, and R. Urtasun, "Sports field localization via deep structured models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5212–5220.

[7] J. Chen and J. J. Little, "Sports camera calibration via synthetic data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[8] L. Sha, J. Hobbs, P. Felsen, X. Wei, P. Lucey, and S. Ganguly, "End-to-end camera calibration for broadcast videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 627–13 636.

[9] X. Nie, S. Chen, and R. Hamid, "A robust and efficient framework for sports-field registration," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1936–1944.

[10] Y.-J. Chu, J.-W. Su, K.-W. Hsiao, C.-Y. Lien, S.-H. Fan, M.-C. Hu, R.-R. Lee, C.-Y. Yao, and H.-K. Chu, "Sports field registration via keypoints-aware label condition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3523–3530.

[11] W. Jiang, J. C. G. Higuera, B. Angles, W. Sun, M. Javan, and K. M. Yi, "Optimizing through learned errors for accurate sports field registration," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 201–210.

[12] F. Shi, P. Marchwica, J. C. G. Higuera, M. Jamieson, M. Javan, and P. Siva, "Self-supervised shape alignment for sports field registration," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 287–296.

[13] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2918–2928.