# 講演会の中継における自動カメラスイッチングのための聴衆の注視オブジェクト推定システム

村上雄亮[†1]　松村耕平[†1]　野間春生[†1]

**概要**：動画配信サイトの急速な普及により，講演会や学会等をインターネットを用いて中継することが一般的になりつつある．講演会や学会を中継する際は講演内容を視聴者に適切に伝えるために，複数のカメラを用い映像を切り替えながら行っている．中継業務については，複数人の専門的な知識を有する技術者が協力して行っており技術者，引いては講演会運営者の負担となっている．本論文ではこの問題を解決するために自動カメラスイッチングのための聴衆の注視オブジェクト推定システムを提案する．

## 1. はじめに

Youtube をはじめとする様々な動画配信サイトが発達してきている．これらの動画配信サイトの発達により放送局を持たないユーザが容易に動画配信，生放送を行うことが可能となっている．上記に加えて動画というメディア形式が情報伝達の面で優れていることより，講演会や学会，企業の発表会などをインターネットを用いて中継することが一般的になりつつある．これによって時間的・身体的な事情により現地に赴けない人々もインターネットを通じて講演会への参加が可能となった．さらに動画として残しておくことで見逃した講演を講演後に聴講できるという利点もある．これらの利点より，我々はこういった傾向は今後も加速していくものと考える．

通常，講演会や学会を中継する際は視聴者に講演内容を適切に伝えるために複数のカメラを用い，放送するという形式が取られている．複数のカメラ映像を切り替えること（以降，スイッチングと記述する）には視聴者に退屈や不快感を与えることなく，講演内容を効果的に伝える効果がある．しかし，スイッチングを適切に行うことは難しく，スイッチングについての専門的な知識を有する技術者（以降，スイッチャと記述する）を必要とする．講演会や学会など多視点映像を用い生放送を行う場合，スイッチャを含めた動画作成や生中継についての専門的な知識を持った複数の技術者（以降，オペレータと記述する）が協力して業務を行うことが必要となり運営者はこれらの手配を行う為，金銭的・時間的な負担を強いられている．

そこで本研究では，オペレータが行っていた様々な業務を自動化することで，オペレータの負担を軽減し，オペレータ単独での中継業務を可能とする支援システムを提案する．生放送に必要となる業務は大まかに，1)登壇者を撮影するカメラの操作，2)複数のカメラから配信映像を選択するスイッチング操作，3)権利問題による放送可否の判断，4)テロップ等の放送情報の付与，の 4 点である．

本研究では，2)複数のカメラから配信映像を選択するスイッチング操作の自動化に着目する．スイッチングを自動化するためのアプローチとして，講演会場にいる聴衆の注目を利用する．講演中，聴衆はその時点で内容を理解するために最も適切であると考えられる対象について注視し聴講していると考えられる．そのため，聴衆の注目に基づくことで視聴者にとって適切に内容を理解できるスイッチングを自動的に行うことを目指す．

## 2. 関連研究

注目点情報は，ディジタルサイネージや映画館等での広告効果の測定やスポーツ観戦における観衆の注目行動の分析など様々な用途で用いることが期待されており，様々な研究がなされている．

横井らのシステムではあらかじめ撮影された高解像度の講義映像から注目点情報を用い講義ビデオを作成する[1]．注目点推測については複数の視聴者があらかじめ撮影された講義映像に対して注目をしている点をポインティングデバイスによって示すことで行っている．この推測結果を用いてトリミングを行い講義ビデオを作成する．

Zhang らの研究は，カメラ前の人物が設定した物体を注視しているかをリアルタイムで識別することが可能であると示している[2]．カメラから取得した顔画像から得られた特徴点を CNN モデルに通すことで注視位置を予測している[3]．この結果を用いることで識別器の学習及び識別を行っている．

上記以外にも頭部カメラを用い集団に属する複数人が同時に注目している点を推測する研究[4]，スマートフォンのインカメラを用いて画面上のどの位置を注視しているか推測する研究[5]などがなされている．

本システムでの注視オブジェクトの推定の対象は講演会や学会の会場の聴衆であるため不特定多数の人物に対応する必要がある．また，このシステムを導入することで聴衆が自然な状態で講演会に参加できなくなることを防ぐため

†1 立命館大学

本システムでは聴衆に無意識な状態で注視オブジェクト推定を行う必要がある．そこで我々は上記の条件に基づき検討を行い，聴衆に向けて設置されたカメラの映像から聴衆の頭部姿勢を推定することで注視オブジェクト推定を行う手法を提案する．

## 3. システムの実装

本システムは，聴衆の動画から画像を切り出し聴衆の頭部姿勢情報を得る．聴衆の頭部姿勢情報を得る部分は Ruiz らの Dockerface[5]及び Hopenet[6]を利用することで GPU サーバ上に実装を行った．全体のシステム構造について図 1 に示す．

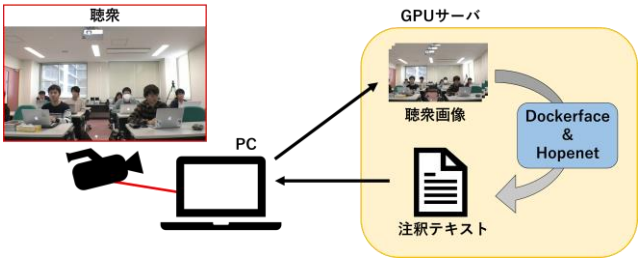


図 1　システムの概要図

### 3.1 頭部姿勢推定

聴衆の様子を撮影した映像から 1 フレーム／秒で画像を切り出し，入力画像として GPU サーバ上の処理システムで処理を行う．処理システムは 2 段階に分かれている．1 段階目では入力画像から顔検出を行い，検出された顔の座標，検出の確かさを示した注釈テキストが出力される．2 段階目では 1 段階目で得られた注釈テキストと入力画像に基づき頭部姿勢推定を行う．頭部姿勢として図 2 に示すような yaw 軸，pitch 軸，roll 軸の 3 軸の角度を推定する．1 段階目と同様に顔の座標，検出の確かさ，頭部の 3 軸の角度を示した注釈テキストを出力する．2 段階目で出力される注釈テキストの内容について図 3 に示す．

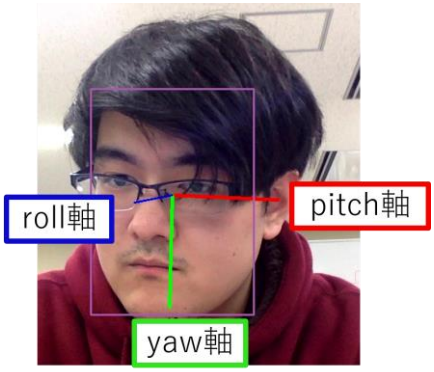


図 2　頭部姿勢として得られる 3 軸の例

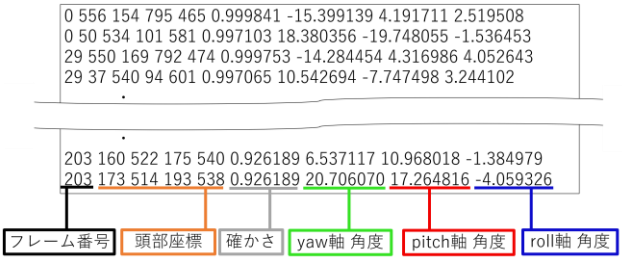


図 3　2 段階目で出力される注釈テキストの内容

### 3.2 動画に対しての処理

3.1 節に示したシステムを利用して WISS2018 の会場に設営した聴衆方向のカメラで撮影された映像について処理を行った．撮影した映像は司会者から聴衆に対してスクリーン，登壇者を注視するように指示を出した 2 種類である．WISS2018 の会場見取り図を図 4 に示す．動画を処理するとそのフレームにおいて検出された顔に対して頭部姿勢推定を行い，出力された yaw 軸の角度について 90° ～-90° を 15 分割したヒストグラムを表示する．スクリーン，登壇者を注視するように指示した映像の処理結果について，それぞれ図 5，図 6 に示す．

注視オブジェクトがスクリーン（図 5）から登壇者（図 6）に変化するに伴ってヒストグラムのピークが 90° 方向に移動することが期待されたが，登壇者を注視オブジェクトに定めた際もヒストグラムのピークは変わらず，そのような結果は得られなかった．期待していた結果が得られなかった理由としては，次の 3 点が挙げられる．

- 頭部姿勢推定精度
  本研究では頭部姿勢推定を行うために，顔画像を用いた Ruiz らの手法を用いた．この場合，顔が極端に傾いている場合や顔の解像度が低い場合，頭部姿勢推定の精度が低下する可能性がある．この問題を避けるためには Zhang らの手法のようにオブジェクト単位での注目の有無を判定する方法が考えられる．
- 聴衆と注目オブジェクトの距離
  聴衆がオブジェクトを注目する際の頭部姿勢（yaw 軸の角度）$\theta$ は聴衆からオブジェクトまでの距離に関係する．オブジェクトから近い位置に存在する聴衆は注目点を変える際に $\theta$ が大きくなるのに対して，オブジェクトから遠い位置に存在する聴衆は $\theta$ が小さくなると考えられる．この問題を避けるためにはカメラの設置位置を検討することや顔認識の際に得られる矩形の大きさからオブジェクトと聴衆の距離を推定し，対象を制限するなどの対応が考えられる．
- 聴衆の分散
  本システムの対象が講演会や学会であるため，聴衆が x 座標方向に分散していることが考えられる．そ

のため今回用いた手法で得られる角度という指標では同一のオブジェクトを注視する場合でも特徴が出ない可能性がある．この問題を避けるためには，システム動作前に，注視対象となる各オブジェクトに対してキャリブレーションを行い，x 座標に応じた修正を行うことが考えられる．

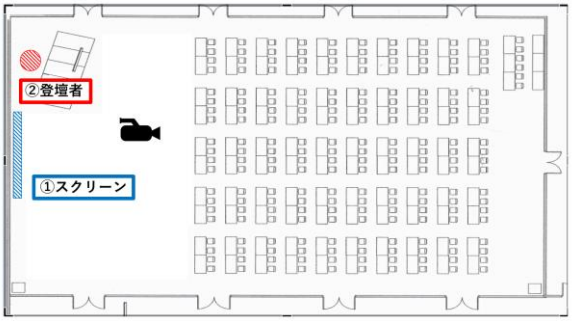


図 4　WISS2018 の会場見取り図

図 5　スクリーンを注視するように指示した
映像の処理結果

図 6　登壇者を注視するように指示した映像の処理結果

## 4. まとめ

本研究では講演会や学会の中継で利用する自動スイッチングシステムのための注視オブジェクト推定システムを提案した．対象を講演会の中継とすることで求められるリアルタイムに不特定多数の人物を対象として意識させることなく注視オブジェクトを推定する手法を検討し，実装した．動画に対して処理を行った結果，注視オブジェクト推定について精度が低いという問題点が得られた．

今後は本研究によって得られた問題点について検討，改善を行い，自動スイッチングシステムへ実装を行う．

## 参考文献

[1] 篠木 雄大，藤吉 弘亘，高解像度映像からの視聴者の注目点を考慮した講義映像の自動生成，映像情報メディア学会誌，2008.
[2] Zhang, X., Sugano, Y., Bulling, A.: “Everyday eye contact detection using unsupervised gaze target discovery,” In: 30th Annual Symposium on User Interface Software and Technology. ACM, 2017
[3] Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: “It’s written all over your face: Fullface appearance-based gaze estimation,” In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017
[4] H.S. Park and J. Shi, “Social saliency prediction,” IEEE Conf. on Computer Vision and Pattern Recognition, pp.4777–4785, June 2015.
[5] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, “Eye tracking for everyone,” Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, pp.2176–2184, June 2016.
[6] N. Ruiz and J. M. Rehg. “Dockerface: an easy to install and use faster r-cnn face detector in a docker container,” arXiv preprint arXiv:1708.04370, 2017.
[7] N. Ruiz and E. Chong and J. M. Rehg, “Fine-Grained Head Pose Estimation Without Keypoints,” arXiv preprint arXiv: 1710.00925, Aug 2018