講演会のライブ中継の自動化に向けた聴衆の注視対象推定システム

村上雄亮[†1,a)]　松村耕平[†1, b)]　大井翔[†1,c)]　野間春生[†1, d)]

概要：動画配信サービスの急速な発達や動画というメディアの特性により，講演会や学会等をインターネットを用いてライブ中継することが一般的になりつつある．講演会等をライブ中継する際は，視聴者に講演内容を適切に伝えるため複数人の専門的な知識を有する技術者が協力して業務を行なっている．そのため講演会の運営者はこれらの技術者を手配する必要があり，負担になっている．本論文では中継業務のうち，カメラのスイッチング業務に着目し自動化を行う．スイッチングの自動化を行うために，会場前方のカメラ及び，聴衆の PC に内蔵されたカメラを用い，聴衆の注目対象を推定するシステムを提案する．提案システムの聴衆の注視対象推定精度の検証として，講義室での講演会を想定した実験をおこなった．その結果，会場前方のカメラに対して PC に内蔵されたカメラの方が高い推定精度を示すことが分かった．一方で PC に内蔵されたカメラを利用する上での問題点も発見された．

1. はじめに

ネットワークインフラストラクチャやストリーミングメディア技術の急速な進歩により，YouTube*1 をはじめとする様々な動画共有サービスが発達してきている．これらの動画共有サービスの発達により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．これらの環境により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で，講演会や学会等をインターネットを用いてライブ中継することが一般的になりつつある．ライブ中継によって時間的・地理的な制約により現地に赴けない人々もインターネットを通じて聴講が可能となる．これにより講演者は講演内容を幅広い人に発信できる．さらに講演を動画として配信することにより視聴者が見逃した講演や再度，視聴したい講演を時間を問わず視聴できるという利点がある[1]．以上より我々は講演会をライブ中継する傾向は今後も加速していくものと考える．

一方で，ライブ中継を行うには様々な業務が必要となり複数人が協力して行っていることが多い．講演中は壇上でのトラブルや著作権コンテンツの配信停止等，様々な状況への対応が必要であるため，講演会の運営者が映像撮影，制作についての専門的な知識を有するプロの技術者（以下，オペレータと記述する）を手配することや，運営メンバ内の配信業務経験者が行うことが基本となっている．これは運営者にとって金銭的・時間的な負担となっており，講演会をライブ中継することに対する弊害となっている．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目指す．ライブ中継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)テロップ等の放送情報の付与, 4)権利問題による放送可否の判断の 4 点である． 本論文では，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に着目し，その自動化を行う．

スイッチングとは複数のカメラで撮影された異なる視点の映像から中継に用いる最適な映像を判断し，映像を切り替える映像制作技術のことである．視聴者に不快感や退屈を感じさせることなく，映像の内容を効果的に伝える効果があり，複雑な映像編集が行えないライブ中継において重要な位置づけとなっている[2]．実際の講演会のライブ中継においても，視聴者に講演内容を適切に伝えるため複数のカメラを用意し，それらをスイッチングすることが多い．しかし適切なスイッチングを行うことは難しく，プロのスイッチングオペレータに依頼することも少なくない．スイッチングオペレータは，コンテンツの内容理解や講演者の行動理解をすることでスイッチングを行なっている．これらをふまえたうえでスイッチングを行うことは難しい．

本研究では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と考えられる対象に注目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スイッチングを行うことでライブ中継の視聴者にとっても内容を適切に理解できる映像となると考えられる．本論文ではスイッチングを自動化するために用いる聴衆の注目対象推定を行うシステムを提案する．

†1　立命館大学
a)　ymurakami@mxdlab.net
b)　marsumur@acm.org
c)　SHO.OOI@outlook.jp
d)　hanoma@fc.ritsumei.ac.jp
*1　http://www.youtube.com

2. 関連研究

講演の一種である大学等の講義映像の自動生成について，様々な研究がなされている．Liu らの研究ではプロの映像制作者へのインタビュから映像制作についての規則を収集し，仮想ディレクタがこの規則と各カメラのステータスに基づきカメラ操作やスイッチングを行う形で講義の自動録画を行っている[3]．篠木らのシステムでは，あらかじめ撮影された高解像度の講義映像から，視聴者の注目点情報を用い講義ビデオを作成する[4]．注目点情報については複数の視聴者が講義映像に対して，注目して見ている点をポインティングデバイスによって指し示すことで推定を行っている．得られる推定結果に基づき元映像にトリミング処理を行い講義ビデオを作成する．

人がどの点・対象を注視しているかという注目点・注目対象情報は，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の用途で用いられることが期待されており，様々な研究がなされている．Zhang らの研究では，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[5]．このモデルを利用することでカメラ前の単独の人物が事前に設定された特定の対象を注視しているかを瞬時に識別することが可能であると示している[6]．Park らの研究では，頭部に装着した一人称視点カメラを用いてイベントなど，複数の社会的集団が存在する場合に人々が注視していると考えられる位置をヒートマップで示すものである[7]．上記以外にも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[8]などがなされている．

既存研究ではスイッチングを行うカメラを決定する際，講演中に生じるイベントに基づき選択を行っている．この方法ではスイッチングが単調なものとなり視聴者が退屈に感じる恐れがある．また，視聴者が見たいと思う対象がイベントの発生していないものであった場合，視聴者は見たいものが見れないといった不快感を感じる可能性がある．そこで我々は会場の聴衆が注目している対象にスイッチングするというアプローチをとることで既存研究と比べて，より視聴者が見たいと思う対象にスイッチングできるシステムを提案する．

3. アプローチ

本研究ではコンテンツの内容理解や講演者の行動理解といったスイッチングの判断要素を，聴衆の注目対象を推定することによって補い，カメラのスイッチングの自動化を図る．注目対象を推定する対象として講演会の会場に訪れている不特定多数の聴衆を設定する．また，本システムはあくまでライブ中継を支援することが目的であり，会場にいる聴衆の聴講を妨げるものであってはならない．よって本システムを導入するにあたり，聴衆に負荷がかかることは避ける必要がある．そこで，聴衆の注目対象を推定する方法として下記の 2 条件に基づき検討を行った．(1) 不特定多数の人物を対象とする．(2) 対象に対して負荷をかけず注目対象の推定を行う．一般的に，人の注目を計測する際に用いられるアイトラッカ等の視線測定装置は聴衆各個人が装着する必要があり，聴衆に負荷をかける状態になってしまうことや特殊な機器を導入するためコストの面からも現実的ではない．そのため本研究では会場の前方から聴衆の方向に設置したカメラで撮影された聴衆の映像をもとに，聴衆の注目対象の推定を行う．

聴衆の映像から注目対象の推定を行う方法として大きく 2 つのタイプがあると考えられる．1 つ目は映像全体に処理を行う方法である．肌色領域面積やオプティカルフローを用いることで推定を行う．この方法では大局的な特徴量を利用するため正確性に欠ける可能性はあるが，計算量を抑えることができる．もう 1 つは映像から聴衆の検出を行い，各聴衆に対して処理を行うことで得られる頭部姿勢や視線情報を用いることで聴衆の注目対象の推定を行う方法である．この方法では映像のフレームから聴衆を検出し、その聴衆に対して処理を行うという 2 段階の処理を含むため計算量は増える．しかし，聴衆各個人に対して処理をかけるため正確性の高い結果が得られると考えられる．本研究では聴衆の注目対象を推定するにあたって，より正確な結果が得られると考えられる後者の方法を用いる．その中でも頭部姿勢の情報を用いることで比較的計算量を抑えた上で高い正確性を保ち聴衆の注目対象推定を行う．

ある情報系の学会（200 名程度）の会場前方のカメラ映像を利用し，聴衆の注目対象を推定する試行について行ったところ，考慮すべき問題が存在することがわかった．問題点として，(1)会場前方からカメラで撮影するため会場後方に存在する聴衆の頭部姿勢について高い精度で得ることができない．(2)会場後方の聴衆の頭部が前方の聴衆に隠されることがあり，それによって注目対象推定の判定材料が減ることになる．の２つが発見できた．

上記の問題による注目対象の推定精度低下について改善するため，会場前方のカメラに加えて聴衆のラップトップ PC に内蔵されたカメラ（以下，PC カメラと記述する）を利用する．PC カメラを利用する理由としては，講演会において聴衆が聴講中に PC を利用することが多くなってきており，特に情報系の学会についてはその傾向が強いと考えられるためである．さらに PC カメラの場合，聴衆各個人の頭部を大きく撮影することができ会場前方のカメラに比べて，個人の頭部姿勢をより高精度に得ることができるためである．

4. 注視対象推定システムの実装

本システムは，聴衆の頭部姿勢を推定するプロセス（4.1），得られた頭部姿勢について注目点に変換するプロセス（4.2），聴衆の注目対象を推定するプロセス（4.3）の 3 つのプロセスから成る．全体のシステム構造について図 1 に示す．また各プロセスについて以下の節で詳しく述べる．

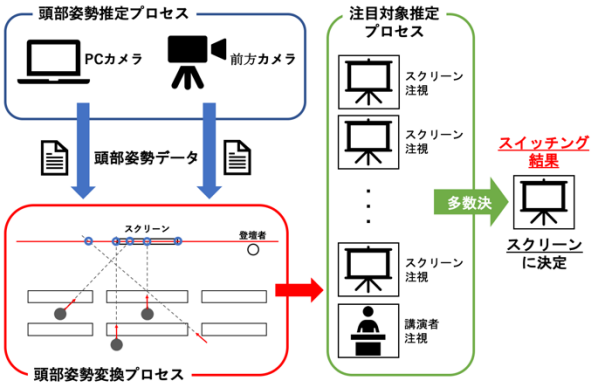

図 1　システムの概要図

4.1 頭部姿勢推定

聴衆の顔検出については Adam の face_recognition[9]を，頭部姿勢推定については Ruiz らの Hopenet[10]を利用することで実装を行った．本プロセスでは聴衆を撮影した映像から 1 フレーム／秒で画像を切り出したものを入力画像として処理を行う．Face_recognition では画像内の顔の検出を行い，顔の座標を出力とする．Hopenet では入力画像と顔の座標を基に頭部姿勢推定を行う．頭部姿勢として図 2 に示すような左右方向の yaw 軸，上下方向の pitch 軸，傾げる方向の roll 軸の 3 軸の角度を出力として推定する．本システムでは yaw 軸角度について，正面を向いている場合を 0°とし右向き 90°，左向き-90°の 180 度で推定を行い，注目対象推定の判定材料とする．

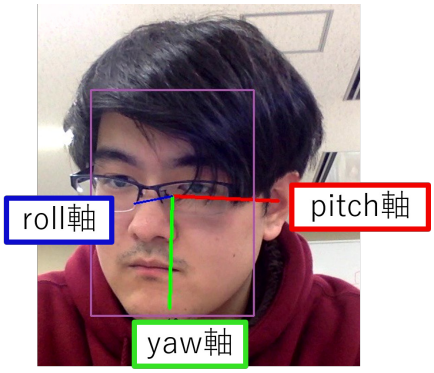

図 2　頭部姿勢として得られる 3 軸

4.2 頭部姿勢の変換

本システムの導入対象としている講演会は様々な会場で開催されている．会場によってサイズや座席の位置は様々であり，それぞれの座席から注目対象への位置関係は異なる．そのため，座席の位置による頭部姿勢と注目対象のぶれを補正する必要がある．各座席からスクリーンの距離と頭部姿勢推定プロセスによって得られた聴衆の頭部姿勢に基づき，図 3 に示すように会場前方のスクリーンが存在している面と頭部正面の線が交わる点を聴衆の注視点とする．

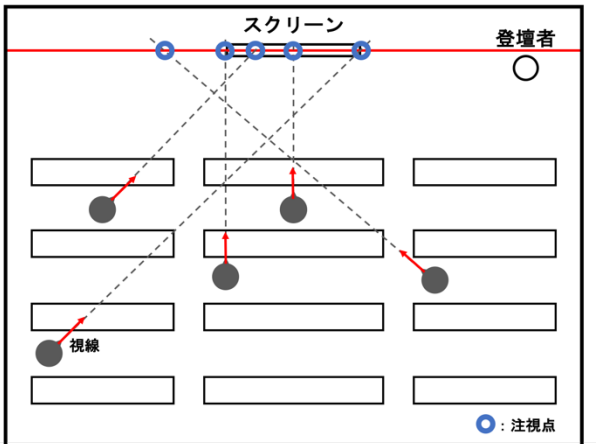

図 3　注視点への変換イメージ

4.3 注目対象の推定

最後のプロセスとして，変換された聴衆の注視点を用いて注目対象を推定する．推定方法として k 近傍法（k-NN）を用いる．k-NN を用いる場合，推定対象となるテストデータとは別に学習データが必要となる．本研究では学習データとして，講演会が開始する前に行うキャリブレーション時のデータを利用する．キャリブレーションではシステム運営者から聴衆に対して，注目対象となる対象物に注視するように指示する．そのようにして撮影された映像に対して前述のプロセスを実行し得られた注目点を学習データとする．学習データに基づき，講演中の聴衆の注目対象を推定する．

また推定結果について，より正確な結果を得るために k-NN での推定に加えて，下記の試行を追加で行う．同一フレームに複数人のデータが検出された場合，同一フレーム内の k-NN の推定結果について多数決し，フレーム内の推定結果を置き換える．

5. 注視対象推定実験

本システムの聴衆の注視対象推定精度の検証として，講義室での講演会を想定した実験を行った．実験は実験者から指示された対象物を注視している際の被験者の映像を講義室前方，及び被験者の PC カメラを用いて撮影するというものである．撮影された映像に対して本システムを用い注目対象推定処理をかけ，指示内容に対して提案システムの推定結果がどの程度一致しているかを測定する．

本実験では講義室前方の 11×4 列の全 44 座席を利用し，講義室前方に設置されているスクリーンとディスプレイを注目対象として定めた．また，前方カメラとして座席の各島の前方に 1 台ずつ，合計 3 台のカメラを設置した．会場の配置図を図 4 に示す．

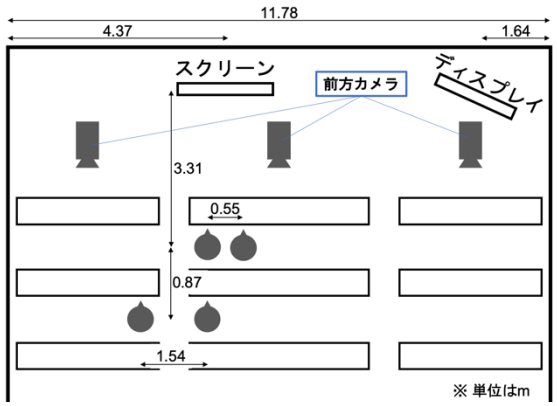

図 4　会場の配置図

指示内容については図 5 に示すように「自身の PC のスクリーン」,「会場前方のスクリーン」,「会場前方のディスプレイ」を各 10 秒間注視する指示を 2 度繰り返すというものを 1 セットとし，合計 30 セット行なった．実験に参加した被験者は男性 5 名である．被験者の座席については毎セット変更し，実験全体を通して各座席に少なくとも 2 人の被験者が座るようにした．

本実験では「会場前方のスクリーン」,「会場前方のディスプレイ」を注視するように指示した時間に得られたデータについて注目対象の推定を行う．また，被験者が頭部を動かす時間を考慮し, 各 10 秒間の前後 2 秒間を除いた 6 秒間のデータを利用した．以上の条件を満たし，被験者の頭部が検出されたデータ数を表 1 に示す．

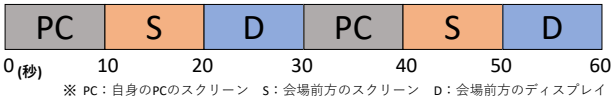

図 5　1 セットの指示内容

表 1　k-NN に利用するデータ数

前方カメラ		被験者のPCカメラ	
学習データ	テストデータ	学習データ	テストデータ
82	1942	66	2067

6. 結果と考察

被験者の注目対象について，前方カメラと PC カメラの各カメラで撮影された映像を用いた時の本システムの推定精度を確認した．各カメラについて k-NN のみを利用した場合の推定精度，及び k-NN の結果を同一フレーム内で多数決した場合の推定精度について表 2，表 3 に示す．

表 2　前方カメラ映像を利用した場合の精度

	k = 2	k = 3	k = 3
k-NN	73 % (1422)	73 % (1422)	70 % (1368)
k-NN + 多数決	80 % (1544)	79 % (1530)	77 % (1499)

※ ( )内は指示内容と推定結果が一致したデータ数

表 3　PC カメラ映像を利用した場合の精度

	k = 2	k = 3	k = 3
k-NN	94 % (1945)	94 % (1940)	94 % (1942)
k-NN + 多数決	97 % (2015)	97 % (2007)	97 % (2013)

※ ( )内は指示内容と推定結果が一致したデータ数

前方カメラと PC カメラの結果を比較すると，前方カメラを用いた場合の精度に対して，被験者の頭部がより鮮明に映っている PC カメラを用いた場合の方が精度が高いことが分かる．また，いずれの結果からも見て取れるが，k-NN のみを利用するのではなく k-NN の推定結果を用いて多数決を行うことで推定精度が向上することが示された．

PC カメラについて，前方カメラの台数より多い 5 台のカメラを利用しているにも関わらず，学習データ数が減少した．この事象の原因として被験者が前方の注目対象を注視する場合, PC カメラ映像には仰ぎ見る角度の被験者が写り，顔検出が出来ない場合があることが考えられる．そのため, PC カメラはあくまで前方カメラのサポートとして用いることが前提となる．

7. まとめと今後の展望

本研究では，講演会のライブ中継におけるスイッチング業務について自動化するための聴衆の注目対象推定システムを提案した．また検証実験を行い，本システムにおける聴衆の注目対象の推定精度を確認した．前方カメラと PC カメラを利用する場合，精度に差が見られることが分かった．両カメラの推定結果について複合したものを利用することが重要であることが示された．

今後は本実験で既知のものとされていた聴衆の座席位置について，前方カメラと PC カメラを複合的に用いることで検出を行い，聴衆の頭部姿勢を注目点に変換する方法について検討する必要がある．また，本システムで推定される聴衆の注目対象に基づき，自動的にスイッチング行うアルゴリズムを作成する．

参考文献

[1] “インターネットを利用した研究発表のライブ動画中継の試行について.”
http://www.sigmus.jp/?page_id=966, (参照 2019-12-18)
[2] NHK 放送技術局. テレビ番組の制作技術 増補版. 兼六館出

版, 2011. 第 3 章 スイッチング, pages 89-140.
[3] Q. Liu, Y. Rui, A. Gupta, and J. J. Cadiz. “Automating camera management for lecture room environments.” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’01, pages 442–449, New York, NY, USA, 2001. ACM.
[4] 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.
[5] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. “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.
[6] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. “Everyday eye contact detection using unsupervised gaze target discovery.” In 30th Annual Symposium on User Interface Software and Technology, ACM, 2017.
[7] H.S. Park, and J. Shi. “Social saliency prediction.” IEEE Conf. on Computer Vision and Pattern Recognition, pp.4777–4785, June 2015.
[8] 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.
[9] A. Geitgey. “Face Recognition.” https://face-recognition.readthedocs.io/en/latest/index.html, (参照 2019-12-18)
[10] N. Ruiz, E. Chong, J. M. Rehg. “Fine-Grained Head Pose Estimation Without Keypoints.” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 2074-2083.