Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 3 (2015) 518 – 525

Procedia MANUFACTURING

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Gesture recognition using a depth camera for human robot collaboration on assembly line

Eva Coupeté*, Fabien Moutarde, Sotiris Manitsaris

Mines ParisTech, Robotics Lab – CAOR, 60 boulevard Saint-Michel, 75006 Paris, France

Abstract

We present a framework and preliminary experimental results for technical gestures recognition using a RGB-D camera. We have studied a collaborative task between a robot and an operator: the assembly of a motor hoses. The goal is to enable the robot to understand which task has just been executed by a human operator in order to anticipate on his actions, to adapt his speed and react properly if an unusual event occurs. The depth camera is placed above the operator, to minimize the possible occlusion on an assembly line, and we track the head and the hands of the operator using the geodesic distance between the head and the pixels of his torso. To describe his movements we used the shape of the shortest routes joining the head and the hands. We then used a discreet HMM to learn and recognize five gestures performed during the motor hoses assembly. By using gesture from the same operator for the learning and the recognition, we reach a good recognition rate of 93%. These results are encouraging and ongoing work will lead us to experiment our set up on a larger pool of operators and recognize the gesture in real time.

Keywords: Human-robot collaboration; Depth camera; Gesture recognition; Hands tracking

1. Introduction

Nowadays, the use of collaborative robots on assembly line is an option which is increasingly explored. These robots would allow an industrial automation of the factories and would improve the productivity in the industrial production plants. On certain workstations, sharing activities between an operator and a robot may be made by

* Corresponding author. Tel.: +33 1 40 51 94 36.
E-mail address: eva.coupete@mines-paristech.fr

doi:10.1016/j.promfg.2015.07.216

letting the low added value tasks or the tasks source of musculoskeletal disorders to the robot. Also, to enable operators to work in confidence and in a safe environment we must equip robots with intelligence so that they understand the events that occur around them, including the actions carried out by the operator.

We are therefore interested by the recognition of technical gestures performed by the operator to enable the robot to understand which task has just been executed in order to anticipate on his actions, to adapt his speed and react properly if an unusual event occurs. These abilities should be sufficient to ensure the safety of the operator nearby.

For his purpose, we used a depth camera, a Kinect[1], and worked with the depth maps. The camera is positioned high filming with a view from above the operator. In this configuration we minimize the occlusions due to the flow of parts on the assembly line and the camera is not obstructing the operator while he is working. Recordings were performed in experimental cells at PSA.

To recognize the gesture, our first work has been to track the hands and the head of the operator. To do this we calculate the geodesic distances of each point of the operator torso to the head center using the Dijkstra algorithm.

For the learning and the recognition we use a combination of K-Means and HMM.

We studied a case where a dual arm robot and an operator cooperate to assemble a motor hoses. We tested our method on offline recognition of 5 gestures and we have 93% of good recognition. The data used for the learning and the recognition are gestures performed by the same operator.

2. Related work

2.1. Human-robot collaboration

With the increasing apparition of robot in our everyday life, the research on human-robot collaboration and interaction has been very active these past years.

Social robots have found numerous applications in human - robot interaction. These robot are used to interact with elderly people [1], or to guide visitors in museum [2] for example. On other cases, the robot and the human are working together on a collaborative task: for example, in [3] the robot helps patient in walk training, and in [4] the robot and the human are carrying the same object for an industrial application.

In the industrial context, the collaboration between a robot and an operator represents a problem on security aspect [5] and on the operator acceptability to work with a new partner [6]. To comply with those needs, new industrial robots are designed to be safer and to provide complementary skill to human co-workers like the Kuka LWR [7] and the Universal Robot UR5 [8].

2.2. Human pose estimation

Human pose estimation using vision is a field which has been well studied these past years due to the large apparition of cameras, and more recently the RGB-D cameras, on numerous supports: laptop, video surveillance, smartphone... Many applications use human poses estimation, like gaming or detecting a mass stampede.

To determine the human pose, we can separate the methods in two classes: using learning or without prior knowledge.

The first class enables systems to be very reliable with a computation time very low after the learning. But they are limited to a number of learning poses. Sun et al. [9] use 3D voxel to recover a body pose, while in [10] the skeleton is reconstructed with a depth map and randomized decision forests. Migniot et al. in [11] use particular filtering with a top view depth camera to determine the position of a top human body.

The second class depends strongly of the image used to estimate the pose, and the computation time is longer than for the first class, but there is no limitation on poses. Schwarz et al. in [12] extract a skeleton with the geodesic

[1] http://www.microsoft.com/en-us/kinectforwindows/

distances, but only with the user facing a RGB-D camera. In [13] multi-cameras are set up and reconstruct 3D poses estimation.

2.3. Gesture recognition

Many methods using vision manage recognition of gestures without estimation of the human pose. For example cloud of space-time interest points can be used for features [14]. The authors in [15] extended to 3D the 2D Harris features [16]. The authors of [17] used Gabor filters to extract features from a video.

The use of a pose estimation is widely used when the goal is to recognize a hand gesture, like in [18] to recognize signs language, and also in [19].

For the recognition of gestures from the whole human body, the Kinect SDK who gives the skeleton of the user, is often used: see for example [20], [21] and [22]. But this SDK works only for poses where the user is facing the camera.

Various approaches have been proposed to handle dynamic gesture recognition. The most known are the HMM (Hidden Markov Model) [23] used for example in [24] and [25]. In [25], a discrete HMM is used to recognize gestures using their velocity. SVM (Support Vector Machine) are also becoming a popular way for visual spatio-temporal pattern recognition as in [26]. Another approach relies on matching with temporal templates, as in [27]. Finally, in [22] the authors use a DBMM (Dynamic Bayesian Mixture Model) to combine multiple classifier likelihoods.

3. Methodology

In this section we present our use case and we explain our methodology to recognize gestures using a depth camera with a top view. We divided our methodology in three steps: the hands tracking, the features extraction and finally the recognition.

3.1. Presentation of the use case

Our use case presents an example of human-robot collaboration in an industrial context. It is inspired by a concrete task, the preparation of motor hoses. These motor hoses are composed of three subparts.

In our set up, the operator is facing a robot, a preparation table is separating them, see Fig. 1(a). First, the robot gives to the operator the two first components (gestures 1 and 2), the operator joins these parts (gesture 3) and screws them together (gesture 4). Then, the robot gives the last component, the operator joins this new part with the first two already assembled and screws a second time. Finally, the operator places the motor horses in a box next to him (gesture 5). To capture the motion of the operator, we use a RGB-D camera, placed above the scene to have a top view, see Fig. 1(b-d). A top view allows minimizing occultation due to all the engines present on an assembly line around this working area.

Fig. 1. (a) set up of our use case, (b) position of the depth camera, (c) RGB image, (d) depth map obtained.

3.2. Hands tracking

To extract the hands location, we used the geodesic distance from the head. We made the assumption that, with a top view, the hands are the visible body parts most remote from the head. An overview of this method is illustrated on Fig. 3.

First, we isolated the torso of the worker by a thresholding of the depth map. Then, we determine the head location by a second thresholding.

All the points that belong to the torso are subsampled by a factor of two to decrease the calculation time. We calculated the geodesic distance using the Dijkstra algorithm [28].

To do this, we connect each point of the subsampled torso with his 8 neighbors, if these neighbors belong also to the subsampled torso. For each of these connections $(I_{xy}, I_{x'y'})$, with I the depth map and I_{xy} the pixel at the location (x, y), we associate a weight equal to the absolute depth difference of the two pixels. The depth is equal to the value of the pixel in the depth map, so $W_{x,y,x',y'} = |I_{xy} - I_{x'y'}|$. To avoid the connection between body parts which are not relevant, for example the head with an arm, which are very often side by side with a top view image, we do not connect two pixels with a weight value above a threshold. We chose a threshold of 50 because it is equal to the half depth difference between the top of the head and the shoulders.

The Dijkstra algorithm calculates, for each point of the subsampled torso, the shortest route between the point and the head. The shortest route between a point A and a point B corresponds at the route which minimizes the sum of the weights associated to connection used to connect the point A to the point B. To improve the calculation time, the shortest route is calculated iteratively, see Fig. 2.

The geodesic distance associated to the point of the torso is finally the sum of the connection weights of the shortest route between the head and the point.

We localize the hand by thresholding the 10% farthest points from the head. If there are more than two blobs, the keep the ones closest to the previous hands location. If there is no previous hand location, we keep the two blobs with the bigger areas.

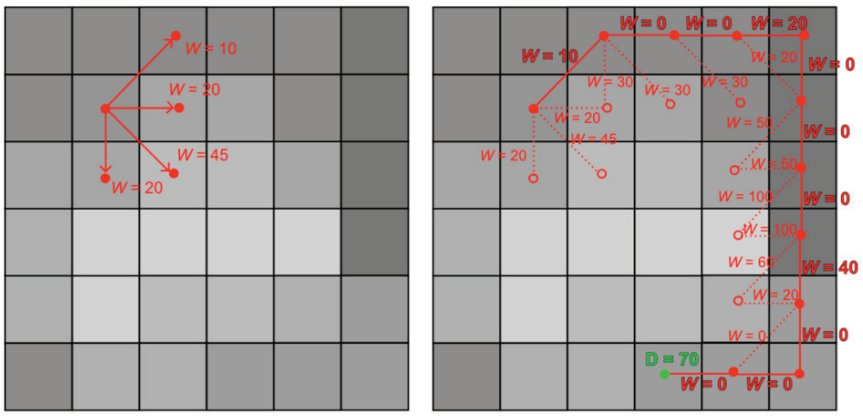

Fig. 2. (a) Weights between neighbour pixels, (b) Calculation of the shortest route between two pixels. The geodesic distance between these two pixels is 70.

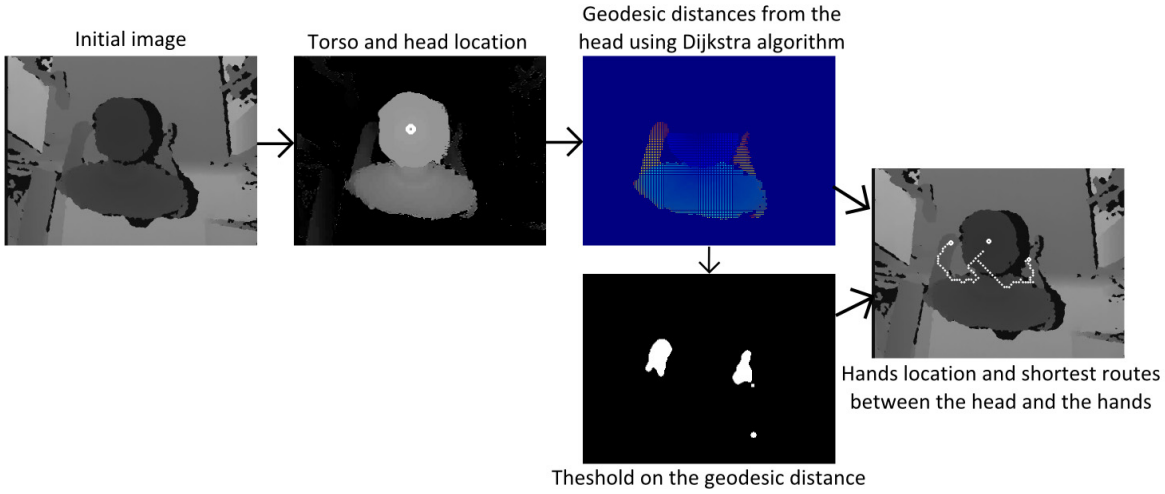

Fig. 3. Overview of the methodology to determine the hand position.

On Fig. 3 we can see in the top left the initial depth map (a). After thresholding, we extract the torso and the head location (b). The next step is the calculation of the geodesic distance between the head and the point of the torso. We can see on the corresponding image (c) that the hands are the most remote parts of the body (red) from the head. We select the ten percent pixels which have the highest geodesic distance (d). We then deduct the hands positions and the shortest routes between these two points and the head (e).

3.3. Features extraction

To describe the gestures done by the operator, we need to have features characterizing these gestures. Since the camera does not move, we do not have to develop a feature that is invariant to rotation and to scale change. We decided to use the shortest route between the hands and the head. These shortest routes evolve with the posture of the operator and can be seen as an approximation of the arms skeletons. We subsampled these routes in 6 samples each, see Fig. 4. We then calculate the 3D vectors between the head and these samples, (1). For each sample, we then have the vector following:

$$V_{head,sample_i} = \begin{pmatrix} x_{sample} - x_{head} \\ y_{sample} - y_{head} \\ I_{sample} - I_{head} \end{pmatrix} \quad (1)$$

With (x_i, y_i) the position of the pixel i and I_i the value of the pixel I in the depth map.

We then concatenate all the vectors in a 36 dimensions vector.

On the Fig. 4 we can see the 12 subsampled points of the shortest routes, 6 for each route. They are represented by 12 red dots. The 3 dimensional vectors are represented by the green arrows.

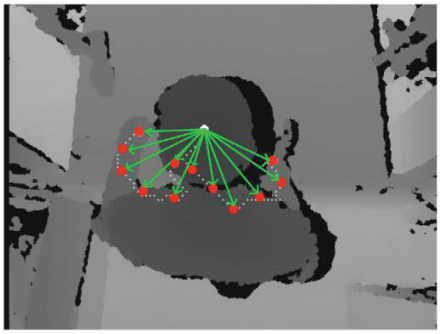

Fig. 4. Vectors head - samples on the shortest route.

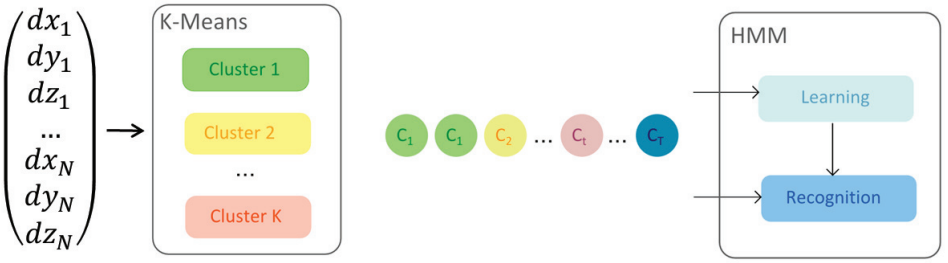

Fig. 5. Overview of the learning and recognition method.

3.4. Learning and recognition

To do the learning and the recognition we used a discreet HMM. We chose to use a HMM because this learning method is suitable to represent temporal phenomenon. In a first step we discretize the features from the learning dataset in K clusters using the K-Means algorithm. The clusterization of the features decreases the calculation time when we are doing the recognition. Indeed, we need to perform real-time recognition in order to obtain a smooth collaboration between the operator and the robot. If we choose a small number of clusters, we lose a lot of information about the gestures. But when we increase the number of cluster, we also increase the calculation time. We set K = 20 which is enough to represent a wide range of postures to discriminate the gestures and enables a real time recognition. Each cluster corresponds to an average position. We then train the HMM with successions of these average positions, see Fig. 5 for an overview.

The learning and recognition were implemented using the GRT (Gesture Recognition Toolkit)[2] library

4. Results

We tested our method to recognize technical gestures. The database for the learning and the recognition consists in gestures performed by the same operator.

We studied five gestures:

- G1 : to take a component in the right claw
- G2 : to take a component in the left claw
- G3 : to join two components

[2] http://www.nickgillian.com/software/grt

- G4 : to screw
- G5 : to put the final motor hoses in a box

To assemble a motor hoses, the gestures 1, 3 and 4 are performed two times each. The gestures 2 and 5 are performed only one time each. This is why we don't have the same number of each gesture in the test database.

The duration of each gesture is between one and two seconds, which is quite short. The recognition has been done off line. We can see the results in the Table 1.

Table 1. Result on gesture recognition with a mono-operator database

		Output Gesture					
		G1	G2	G3	G4	G5	**Recall**
Input Gesture	G1	**66**			1		99%
	G2		**24**	1	3	3	77%
	G3		2	**58**	4	2	88%
	G4				**78**		100%
	G5		3			**32**	91%
	Precision	100%	83%	98%	91%	86%	**93%**

The gestures are well recognized with 93% of good recognition. We can see that the gesture 3 (to join two components) is often mistaken with the gestures 2 (to take a component in the left claw), the gestures 4 (to screw) and the gesture 5 (to put the final motor hoses in a box). This can be explain by the fact that the gesture 3 has less distinguishing features than the other gestures, indeed the operator is more static when he is doing this gesture that when he is doing the other gestures. We can also observe that the gesture 2 is mistaken with the gestures 4 and 5. This is because while the operator is doing these three gestures, he turns on the left side which can be confusing.

5. Conclusion

We have shown a method to recognize technical gestures with a top view by a depth camera. We explained how to extract features using geodesic distance with this set up and we tested our recognition method with success with 93% of the gestures recognized.

But, to make this method suitable to an industrial environment we need to make it work with a large range of operators. We recorded gestures from a larger pool of operators and we are working on gestures recognition with datasets from different persons for the learning and the recognition. After that, we will work on real-time recognition. We also plan to add inertial sensors on operator gloves or on the tools used during the assembly task to discriminate gesture if the vision is not reliable enough.

Acknowledgements

This research benefited from the support of the Chair 'PSA Peugeot Citroën - Robotics and Virtual Reality', led by MINES ParisTech and supported by PEUGEOT S.A. The partners of the Chair cannot be held accountable for the content of this paper, which engages the authors' responsibility only.

References

[1] M. L. Walters, K. L. Koay, D. S. Syrdal, A. Campbell, and K. Dautenhahn, "Companion robots for elderly people: Using theatre to investigate potential users' views", Proc. 22nd IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN 2013), Gyeongju (Korea), pp. 691–696, 26-29 Aug. 2013.

[2] S. Thrun and M. Bennewitz, "MINERVA: A second-generation museum tour-guide robot", Proc. IEEE Int. Conf. on Robotics and Automation, vol. 3, pp.1999-2005, 1999.

[3] T. Sakaki, N. Ushimi, K. Aoki, K. Fujiie, R. Katamoto, A. Sugyo, and Y. Kihara, "Rehabilitation robots assisting in walking training for SCI patient", Proc. 22nd IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN 2013), Gyeongju (Korea), pp. 814–819, 26-29 Aug. 2013.
[4] T. Wojtara, M. Uchihara, H. Murayama, S. Shimoda, S. Sakai, H. Fujimoto, and H. Kimura, "Human-robot collaboration in precise positioning of a three-dimensional object", Automatica, vol. 45, no. 2, pp. 333–342, 2009.
[5] P. Rybski, P. Anderson-Sprecher, D. Huber, C. Niessl, and R. Simmons, "Sensor fusion for human safety in industrial workcells", IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2012), pp. 3612–3619, Oct. 2012.
[6] V. Weistroffer, A. Paljic, P. Fuchs, O. Hugues, J. Chodacki, P. Ligot, A. Morais, and A. H. Collaboration, "Assessing the Acceptability of Human-Robot Co-Presence on Assembly Lines: A Comparison Between Actual Situations and their Virtual Reality Counterparts", Proc. 23rd IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN 2014), Edinburgh (United Kingdom), Aug. 2014.
[7] S. Haddadin, M. Suppa, S. Fuchs, T. Bodenmüller, A. Albu-Schäffer, and G. Hirzinger, "Towards the robotic co-worker", Springer Tracts on Advanced Robotics, vol. 70, pp. 261–282, 2011.
[8] J. Schrimpf, M. Lind, and G. Mathisen, "Real-Time Analysis of a Multi-Robot Sewing Cell", IEEE Int. Conf. on Industrial Technology (ICIT), pp. 163–168, 25-28 Feb. 2013.
[9] Y. Sun, M. Bray, A. Thayananthan, B. Yuan, and P. Torr, "Regression-Based Human Motion Capture From Voxel Data", Proc. British Machine Vision Conference (BMVC 2006), pp. 277-286, 4-7 Sept. 2006.
[10] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images", Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2011), pp. 1297–1304, June 2011.
[11] C. Migniot and F. Ababsa, "3D Human Tracking from Depth Cue in a Buying Behavior Analysis Context", Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, vol. 8047, pp. 482-489, 2013.
[12] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab, "Human skeleton tracking from depth data using geodesic distances and optical flow", Image and Vision Computing, vol. 30, issue 3, pp. 217–226, 2012.
[13] R. Kehl and L. Van Gool, "Markerless tracking of complex human motions from multiple views", Computer Vision and Image Understanding, vol. 104, issue 2–3, pp. 190–209, 2006.
[14] M. Bregonzio, "Recognising action as clouds of space-time interest points", 2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2009), pp. 1948–1955, 20-25 June 2009.
[15] I. Laptev and T. Lindeberg, "Space-time interest points", Proc. 9th IEEE Int. Conf. on Computer Vision (ICCV 2003), vol. 1, pp. 432–439, 2003.
[16] C. Harris and M. Stephens, "A Combined Corner and Edge Detection", 4th Alvey Vision Conference, pp. 147-151, 1988.
[17] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features", 2nd IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72, 15-16 Oct. 2005.
[18] Q. C. Q. Chen, N. D. Georganas, and E. M. Petriu, "Real-time Vision-based Hand Gesture Recognition Using Haar-like Features", Proc. IEEE Conf. on Instrumentation and Measurement Technology, pp. 1-6, 1-3 May 2007.
[19] Y. Fang, K. Wang, J. Cheng, and H. Lu, "A Real-Time Hand Gesture Recognition Method", IEEE Int. Conf. on Multimedia Expo, pp. 995–998, 2-5 July 2007.
[20] K. Lai, J. Konrad, and P. Ishwar, "A gesture-driven computer interface using Kinect", IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 185–188, 22-24 Apr. 2012.
[21] O. Patsadu, C. Nukoolkit, and B. Watanapa, "Human gesture recognition using Kinect camera", Int. Joint Conf. on Computer Science and Software Engineering (JCSSE), pp. 28–32, 30 May-1 June 2012.
[22] D. R. Faria, C. Premebida, and U. Nunes, "A Probabilistic Approach for Human Everyday Activities Recognition using Body Motion from RGB-D Images", Proc. 23rd IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN 2014), Edinburgh (United Kingdom), pp. 732-737, 25-29 Aug. 2014.
[23] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, vol. 77, No 2, pp. 257–286, 1989.
[24] L. Nianjun, B. C. Lovell, P. J. Kootsookos, and R. I. A. Davis, "Model Structure Selection and Training Algorithms for an HMM Gesture Recognition System", Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition, pp. 100–105, 26-29 Oct. 2004.
[25] F. G. Hofmann, P. Heyer, and G. Hommel, "Velocity profile based recognition of dynamic gestures with discrete Hidden Markov Models", Gesture and Sign Language in Human-Computer Interaction, Lecture Notes in Computer Science, vol. 1371, pp. 81–95, 1998.
[26] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach", Proc. 17th Int. Conf. on Pattern Recognition (ICPR 2004), vol. 3, pp. 32–36 , 23-26 Aug. 2004.
[27] A. Bobick and J. Davis, "The recognition of human movement using temporal templates", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257–267, 2001.
[28] E. W. Dijkstra, "A note on two problems in connexion with graphs", Numerische Mathematik, vol. 1, no. 1, pp. 269–271, 1959.